Slowness Learning: Mathematical Approaches and Synaptic Mechanisms

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Herr Dipl.-Phys. Henning Sprekeler
geboren am 8.9.1976 in Münster/Westfalen

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Christian Limberg

Gutachter:

1. Prof. Dr. Laurenz Wiskott
2. Dr. Richard Kempter
3. Prof. Dr. Wulfram Gerstner

eingereicht am:	26. März 2008
Tag der mündlichen Prüfung:	14. November 2008

Abstract

In this thesis, we investigate slowness as an unsupervised learning principle of sensory processing. Two aspects are given particular emphasis: (a) the mathematical analysis of Slow Feature Analysis (SFA) as one particular implementation of slowness learning and (b) the question, how slowness learning can be implemented in a biologically plausible fashion.

In the first part of the thesis, we develop a mathematical framework for SFA and show that the optimal functions for SFA are the solutions of a partial differential eigenvalue problem. The theory allows (a) to make analytical predictions for the behavior of complicated applications and (b) an intuitive understanding of how the statistics of the input data are reflected in the optimal functions of SFA. The theory is applied to the learning of place and head-direction representations and to the learning of complex cell receptive fields as found in primary visual cortex. As a technical application, we use the theoretical results to develop and test a new algorithm for nonlinear blind source separation. The first part of the thesis is concluded by an information-theoretic analysis of the relation between slowness learning and predictive coding.

In the second part of the thesis, we study the question, how slowness learning could be implemented in a biologically plausible manner. To this end, we first show that spike timing-dependent plasticity can under certain conditions be interpreted as an implementation of slowness learning. Finally, we show that both gradient-based slowness learning and spike timing-dependent plasticity lead to receptive field dynamics that can be described in terms of reaction-diffusion equations.

Keywords:
Theory of Slow Feature Analysis, Synaptic Plasticity, Nonlinear blind source separation, place and head-direction cells

Zusammenfassung

In dieser Doktorarbeit wird Langsamkeit als unüberwachtes Lernprinzip in sensorischen Systemen untersucht. Dabei wird zwei Aspekten besondere Aufmerksamkeit gewidmet: der mathematischen Analyse von Slow Feature Analysis - einer Implementierung des Langsamkeitsprinzips - und der Frage, wie das Langsamkeitsprinzip biologisch umgesetzt werden kann.

Im ersten Teil wird zunächst eine mathematische Theorie für Slow Feature Analysis entwickelt, die zeigt, dass die optimalen Funktionen für Slow Feature Analysis die Lösungen einer partiellen Differentialgleichung sind. Die Theorie erlaubt, das Verhalten komplizierter Anwendungen analytisch vorherzusagen und intuitiv zu verstehen. Als konkrete Anwendungen wird das Erlernen von Orts- und Kopfrichtungszellen, sowie von komplexen Zellen im primären visuellen Kortex vorgestellt. Im Rahmen einer technischen Anwendung werden die theoretischen Ergebnisse verwendet, um einen neuen Algorithmus für nichtlineare blinde Quellentrennung zu entwickeln und zu testen. Als Abschluss des ersten Teils wird die Beziehung zwischen dem Langsamkeitsprinzip und dem Lernprinzip der verhersagenden Kodierung mit Hilfe eines informationstheoretischen Ansatzes untersucht.

Der zweite Teil der Arbeit befasst sich mit der Frage der biologischen Implementierung des Langsamkeitsprinzips. Dazu wird zunächst gezeigt, dass Spikezeit-abhängige Plastizität unter bestimmten Bedingungen als Implementierung des Langsamkeitsprinzips verstanden werden kann. Abschließend wird gezeigt, dass sich die Lerndynamik sowohl von gradientenbasiertem Langsamkeitslernen als auch von Spikezeit-abhängiger Plastizität mathematisch durch Reaktions-Diffusions-Gleichungen beschreiben lässt.

Schlagwörter:
Slow Feature Analysis, Nichtlineare blinde Quellentrennung, Ortszellen, Kopfrichtungszellen

To my parents

Contents

Chapter 1

Introduction

We see a cup of coffee on the breakfast table in front of us, reach out and grasp it. Having guided it to our mouth, we take a hot sip and set it back on the table. This simple sequence of actions appears so natural to most of us that we rarely notice how complex it actually is. How do we know that this object on the table is a cup of coffee, although the cup may be white, blue or yellow, large or small, close or far away, topped by milk foam or not? Grasping the cup requires that we know its shape. How do we manage to guess its 3-dimensional shape from the 2-dimensional visual images we perceive with our eyes? Once we know the shape of the cup, we have to move our hand to grasp it and hold it securely, continuously taking into account visual, proprioceptive and somatosensory information that tells us where our hand is in relation to the cup, if the cup is too hot, if we have grasped it the way we planned it and if the grasp is sufficiently strong to prevent it from slipping out of our hand. How do we do that?

The simple action of taking a sip of coffee can be divided into a series of apparently simple problems that turn out to be immensely complicated upon closer inspection. Even worse, this complexity is immanent in most of our daily actions. Nevertheless, we perform them with astonishing ease. How is this possible? The answer to this question is that we have an extremely proficient organ that acts as our central information processing unit and coordinator: Our brain. Somehow, it manages to solve most of the problems we encounter in everyday life within the blink of an eye. How exactly it does that, is one of the formidable questions in modern science and, as yet, mostly unanswered.

The task our brain is facing can roughly be divided into three subtasks. Firstly, sensory information transmitted from our primary sensory organs has to be interpreted. In particular, the sensory data needs to be filtered: relevant aspects should be kept, irrelevant aspects discarded. Secondly, interpreted sensory information has to be stored, integrated with previously acquired knowledge about the environment and our current situation and a plan has to be developed about what action should be pursued. Thirdly, this plan needs to be translated into a detailed set of commands that initiate the actual motor action. It should be kept in mind that such a subdivision can at best be a helpful abstraction, because the borders between the subtasks are drawn somewhat arbitrarily. Moreover, the subtasks are probably mutually entangled by an abundance of feedback loops, so that they may not be clearly separable, neither conceptually nor in their biological implementation.

Interpretation of Sensory Signals

The research presented in this thesis revolves around the first subtask, that is, the interpretation of sensory signals. One of the problems our brain has to solve in this context is that the primary sensory signals reflect information about our surrounding in a very inconvenient manner. For example, in human vision, more than a hundred million retinal receptor cells continuously provide signals about how much light each of them perceives in a very small subregion of our visual field. The resulting maelstrom of "pixel values" contains all the information necessary to, for example, recognize an approaching object, but the representation in the form of pixels is so inconvenient, that the computational power of current computer systems is often insufficient to extract the information we are interested in. Our brain performs this extraction with astonishing speed and reliability, yet the strategies it uses remain almost completely elusive. Understanding the mechanisms at work in our sensory systems is an interesting problem per se, but it can also inspire technical applications in computer vision and other signal processing contexts.

One way of studying sensory systems is by asking what determines their structure. Is the structure of sensory cortex a result of evolution in the sense that the genetic code defines each and every neuron down to its morphology, its ion channel composition and its synaptic connections? Such a hard-wiring hypothesis in its extreme form is biologically implausible, simply because the genome is not large enough to store all of the information necessary. It is more likely that evolution has established a basic set of mechanisms that support robust self-organization. The hope of neuroscience – and maybe of life science in general – is that the number of these mechanisms is sufficiently small, allowing reduced (and intelligible) models.

Learning in Sensory Systems

A question of considerable interest is the role of sensory signals in these self-organization processes. This role probably varies, depending, e.g., on the processing level. It is thought for example, that the basic wiring between the retina and the thalamus, the earliest stages of visual processing, is in some animals established even before the animal opens its eyes. Candidate mechanisms are internally generated chemical gradients (McLaughlin & O'Leary, 2005) and spontaneous activity patterns on the retina (see, e.g. Wong, 1999), both of which should be largely independent of sensory signals. For higher visual areas, in contrast, it is more likely that visual information plays a crucial role in shaping neuronal response properties. It is hard to conceive a genetically determined mechanism that generates highly selective neurons responding to faces of specific contemporary movie stars (Quiroga et al., 2005).

Whenever external signals influence the self-organization process and the resulting structure, the animal is effectively adapting to its environment. One fascinating property of the brain is that it remains highly adaptive throughout life. Its ability to change its structure in response to both external and internal signals is commonly referred as *plasticity* and forms the basis of learning and memory. A basic assumption of this thesis is that the ability of our brain to interpret primary sensory signals is at least partly due to an adaptation to the environment, and hence the result of a learning process. But what are the principles governing this learning process?

There are several ways of tackling this question theoretically. One way is to start with physiology, that is, with mechanisms of neuronal plasticity that have been char-

acterized in the laboratory. Building phenomenological, "mechanistic" models of these mechanisms can help to understand, if they are sufficient to explain the phenomena found in sensory systems, e.g., the response behavior of neurons in the visual system. In addition, such models can help in detecting putative "principles of information processing" of which physiological mechanisms are merely an implementation. In the light of the intimidating wealth of known mechanisms for neuronal plasticity, such abstractions will be indispensable for an understanding of their effect on cortical information processing.

A different approach to sensory learning is to start with abstract principles that are motivated by theoretical considerations on information processing in the brain. Several principles have been proposed, examples of which are optimal data compression approaches (e.g., principal component analysis, Joliffe, 1986) or approaches based on predictive (e.g., Rao & Ballard, 1999) and efficient coding strategies (e.g., independent component analysis, Hyvärinen et al., 2001, or information maximization, Brenner et al., 2000). For abstract approaches to be a candidate model for information processing in the brain, two basic questions need to be answered: (a) Is the approach implementable by physiological mechanisms and (b) can it reproduce phenomena found in sensory areas, e.g., the response behavior of cortical neurons?

Overview of the Thesis

In this thesis, we will approach these questions for the abstract learning principle of *slowness*, which has been proposed as a means for learning invariant sensory representations. The motivation of the slowness principle and its implementation in the form of *Slow Feature Analysis* (SFA) is introduced in chapter 2.

The goal of this thesis is two-fold. In Part I, we extend previous analytical work on SFA (Wiskott, 2003) and develop a mathematical framework, which not only allows to make analytical predictions for applications, but also helps to get an intuitive understanding of the mechanisms of SFA. In chapter 3, we present a theoretical description of SFA that is applied to a set of problems in chapter 4. Based on the theory, we propose a new algorithm for nonlinear blind source separation (section 4.1) and present analytical results for the applications of SFA for learning place and head-direction codes in the hippocampal formation of rodents (section 4.2, published in (Franzius, Sprekeler & Wiskott, 2007)). In chapter 5, we present a mathematical analysis of the model for complex cell receptive fields in primary visual cortex as presented by Berkes & Wiskott (2005). Chapter 6 provides an information theoretic link between Slow Feature Analysis and the principle of predictive coding.

In Part II, we analyze, if the slowness principle can be implemented by biologically plausible means. We show that spike-timing-dependent plasticity, a form of synaptic plasticity that has recently been subject to intense research, can under certain conditions be interpreted as an implementation of the slowness principle. Finally, in chapter 8, we give an outlook on a mathematical description of receptive field dynamics based on (a) gradient-based slowness learning and (b) synaptic plasticity. The resulting description is formally closely related to reaction-diffusion systems, which provides a link to the broad field of self-organized pattern formation.

My hope in presenting this research is that it may be of interest to a relatively diverse audience. The mathematical analysis of Slow Feature Analysis may be of interest to readers working on data analysis and machine learning, because it helps to gain an intuitive and mathematical understanding of how the statistical structure of data is

reflected in the output signals obtained by SFA. The intuition we gained from the theory formed the basis for the new algorithm for nonlinear blind source separation and I hope that it will inspire more applications in the future.

The biological applications of SFA, in particular the results for place- and head-direction codes presented in section 4.2, may be of interest not only to computational neuroscientists modeling visual receptive fields or spatial cognition, but also to researchers interested in computer vision or navigation problems in robotics.

For researchers involved in synaptic plasticity, chapter 7 could be of interest. In showing that temporally non-local Hebbian learning can act as a means for learning sensory invariances, we contribute one more item to the growing collection of functional interpretations of these learning rules that will hopefully one day be merged into a "big picture".

Last, but not least, chapter 8 reveals a link between receptive field dynamics and the theory of reaction-diffusion systems. We feel that this chapter may serve as a starting point for a new approach to receptive field dynamics that could provide modelers with new mathematical tools as well as insights from the well-developed theory of pattern formation.

Chapter 2

The Slowness Principle

2.1 Motivation

Basic Forms of Learning

In computational models of learning, the basic setup is invariably the same: A system receives input signals $\mathbf{x}_\mu$ and produces output signals $\mathbf{y}_\mu$ that depend not only on the input data, but also on a set of internal parameters $\mathbf{w}$. The learning process consists of an adaptation of these internal parameters $\mathbf{w}$ to a given set of so-called *training data*.

The theory of learning distinguishes three basic forms of learning: supervised, reinforcement and unsupervised learning. In *supervised learning* the training data consist not only of a set of input data $\mathbf{x}$, but contains additional external teaching signals that tell the system exactly what output $\mathbf{y}$ it should produce to a given input signal. The goal of learning is to adapt the internal parameters in order to reproduce the desired output as accurately as possible. Learning vocabulary would be a typical everyday example of supervised learning. *Reinforcement learning* differs in that there is still some external "observer" that evaluates the performance of the system, but instead of providing the correct output signal, feedback is restricted to a reward or a punishment signal. An everyday example of reinforcement learning would be a child learning to ride a bike. When it falls, it is often difficult to describe exactly what went wrong or what should be changed in order to succeed. Instead, the child learns to avoid a relatively unspecific negative feedback signal, i.e., the pain associated with falling.

At least for early sensory areas, these two forms of learning are probably an inappropriate description. It is unlikely that there is an external signal telling a simple cell in V1 to develop a Gabor-shaped receptive field or punishing it if it fails to develop one. For this reason, most theoretical models for cells in early sensory areas are based on *unsupervised learning*. In unsupervised learning the training data contain no teaching or reward signal. Instead, the system has access to a set of input data and possesses an internal guideline that governs the adaptation to certain, often statistical features of these data. A typical example for unsupervised learning is Principal Component Analysis (PCA). In PCA, the guideline is to compress an input vector into a lower-dimensional representation that minimizes the quadratic reconstruction error. In this case, the internal guideline of the learning process is the minimization of the error. Technically, this is done by extracting input dimensions with high variance. Other examples are *efficient coding* approaches such as Independent Component Analysis (ICA, Hyvärinen et al., 2001) or sparse coding (Olshausen & Field, 2004), which aim at removing statistical redundancies in the input

data.

The common theme in all these techniques is that there is an underlying ad-hoc principle that governs the adaptation to the training data. Note that there may be different implementations of the same principle. An example: The detection of input dimensions with large variance in PCA can be done either by means of an eigenvalue approach or by Hebbian learning with a linear neuron (Oja, 1982).

Invariant Representations

The research presented in this thesis circles around the problem of how to learn invariant sensory representations. There are several reasons why these representations are of interest. For once, it is known that a variety of cell types in the brain displays invariant responses: complex cells in primary visual cortex respond to gratings of a certain orientation, but are invariant to their exact position; place cells in the hippocampal formation of rodents fire whenever the rat is at a particular location, irrespective of the orientation of its head (Muller et al., 1994); cells in the temporal lobe of humans seem to respond selectively to images of certain contemporary movie stars, but are invariant with respect to high level differences such as clothing, posture or perspective (Quiroga et al., 2005). It is likely that these representations are the neural correlate for the remarkable ability of humans to recognize objects or locations in spite of ever changing conditions. This ability is indispensable for successful interaction with the environment, because objects never look exactly the same and a deficit in recognizing them would render them useless to us in everyday life. For the same reasons of robustness, invariant object recognition is relevant for applications in robotics or computer vision as well.

It is by no means clear how such representations are established. Common agreement is, however, that they are at least partially learned from experience. Supervised learning is not a good candidate, because providing the teaching signal would be a rather cumbersome task ("This is a car from the left and this is a car from the right and this is a car from the front and ...". In the worst case, we would have to repeat this process for all possible objects and perspectives). This leaves us with reinforcement and unsupervised learning. It is likely that reinforcement signals acquired through the interaction with the environment play a role in the formation of invariant representations, but a theoretical description of these effects requires relatively complicated models that take behavioral aspects into account. From the perspective of Occam's razor, unsupervised learning is more appealing, because it allows much simpler models. Moreover, it is interesting to examine to what extent invariant representations can be understood as a result of an adaptation to the statistics of sensory data alone.

The Slowness Principle

One approach to the unsupervised learning of invariant representations is based on the observation that behaviorally relevant signals such as the identity of an object typically vary on a much slower time scale than the primary sensory signals they evoke. An illustrative example is the observation of a running zebra. The optical signal perceived by a single receptor cell in the retina will be a quick succession of "black" and "white", corresponding to the stripes of the zebra. From these quickly varying signals, our brain somehow manages to extract the fact that there is a zebra in the scene. In general, once there is a zebra, it remains present for a few seconds at least, so the abstract

representation "zebra" will vary on a much slower time scale than the primary sensory signals.

This observation can serve as a heuristic for the interpretation of primary sensory signals: Slowly varying aspects are more likely to be behaviorally relevant than those that vary quickly. Deriving an unsupervised learning principle from this observation is straight-forward: Given a set of training data, adapt the internal parameters $\mathbf{w}$ such that the resulting output signals vary on a behavioral time scale, or - even simpler - as slowly as possible. This is a simple formulation of the *Slowness Principle*.

In any application of the slowness principle, one caveat must be avoided: One easy way of generating slowly varying output signals is by simply low-pass filtering the input data. It is not likely, however, that the application of a low-pass filter to primary sensory data yields behaviorally relevant information. In addition, low-pass filters require a certain integration time and thus tend to reduce processing speed. This may lead to longer reaction times, which is not desirable for an animal that needs to react quickly to possibly dangerous changes of its environment. The slowness principle can thus only be expected to yield meaningful results, when the temporal integration time of the mapping between input and output is limited.

Although so far we have motivated the slowness principle in terms of behavioral relevance, it is easy to conceive that it is well suited for learning invariant representations. Slowly varying output signals can only be generated if the output signal is independent of the most quickly varying changes in the input data. For example, a system that generates a slowly varying output signal from a video showing a quickly rotating object has to be invariant with respect to the rotation angle of the object, otherwise it would not vary slowly.

The slowness principle forms the basis of a whole class of algorithms for invariance learning (Földiàk, 1991; Mitchison, 1991; Becker & Hinton, 1992; O'Reilly & Johnson, 1994; Stone & Bray, 1995; Wallis & Rolls, 1997; Peng et al., 1998; K. Körding et al., 2004). Here, we will mostly concentrate on one particular implementation of this principle: Slow feature analysis (SFA) as introduced by Wiskott & Sejnowski (2002).

2.2 Slow Feature Analysis

Slow feature analysis is based on the following learning task: Given a multi-dimensional input signal, find scalar input-output functions that generate output signals that vary as slowly as possible but carry significant information. To ensure the latter the output signals are required to be uncorrelated and have unit variance. In mathematical terms, this can be stated as follows:

***Optimization problem 1**: Given a function space $\mathcal{F}$ and an N-dimensional input signal $\boldsymbol{x}(t)$ find a set of J real-valued input-output functions $g_j(\boldsymbol{x})$ such that the output signals $y_j(t) := g_j(\boldsymbol{x}(t))$ minimize*

$$\Delta(y_j) = \langle \dot{y}_j^2 \rangle_t \tag{2.1}$$

under the constraints

$$\langle y_j \rangle_t = 0 \qquad \textit{(zero mean)}, \tag{2.2}$$

$$\langle y_j^2 \rangle_t = 1 \qquad \textit{(unit variance)}, \tag{2.3}$$

$$\forall i < j : \langle y_i y_j \rangle_t = 0 \qquad \textit{(decorrelation and order)}, \tag{2.4}$$

with $\langle \cdot \rangle_t$ and $\dot{y}$ indicating temporal averaging and the derivative of y, respectively.

Equation (2.1) introduces the Δ-value, which is a measure of the temporal slowness (or rather 'fastness') of the signal $y(t)$. The constraints (2.2) and (2.3) avoid the trivial constant solution. Constraint (2.4) ensures that different functions g_j code for different aspects of the input.

It is important to note that although the objective is the slowness of the output signal, the functions g_j are instantaneous functions of the input, so that slowness cannot be achieved by low-pass filtering. Slow output signals can only be obtained if the input signal contains slowly varying features that can be extracted by the functions g_j.

Depending on the dimensionality of the function space $\mathcal{F}$, the solution of the optimization problem requires different techniques. It has been shown that for finite-dimensional function spaces, the problem can be reduced to a (generalized) eigenvalue problem (Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2005). This allows to solve the optimization problem by means of a computationally efficient algorithm, that is introduced in section 2.3. The focus of this thesis will be on the more complicated situation, where the function space is infinite-dimensional. In this case, the solution of the problem can no longer be found by means of linear algebra, but requires variational calculus instead.

In part I of the thesis, we will present mathematical approaches for two particular classes of problems. First, we will consider the case where there are no restrictions on the function space $\mathcal{F}$ and where the set of possible input data can be parametrized by a finite number of parameters. Second, we will develop a mathematical framework for the case where the input data are possibly infinite-dimensional, but generated by a finite number of continuous transformations. The theory for the first scenario is presented in chapter 3 and applied to two concrete problems in chapter 4. The theory for the latter case is developed in chapter 5 and used to derive analytical results for the model of complex cell receptive fields presented by Berkes & Wiskott (2005). In chapter 6, we will provide an information-theoretic link between SFA and predictive coding.

At first glance, the scenarios treated in chapters 3-5 appear rather academic, because neither infinite-dimensional input signals nor infinite-dimensional function spaces can be realized in practice. The reader will see, however, that these "abstract" cases provide mathematical tools that are unavailable for the finite-dimensional case. The resulting mathematical framework allows to make analytical predictions for systems that could previously only be simulated. The agreement of the predictions with simulation results shows that the infinite-dimensional cases provide a decent description of real applications of SFA. Moreover, they allow to develop an intuitive understanding of how the structure of the input data is reflected in the functions that are optimal for SFA.

2.3 The SFA Algorithm

As mentioned above, the optimization problem for SFA can be solved efficiently by means of a generalized eigenvalue problem. To sketch how this is done, let us assume that the function space $\mathcal{F}$ is finite-dimensional with dimension M and that it is spanned by a set of basis functions f_i. For reasons of notational compactness, let us arrange the functions f_i in a vector-valued function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_M(\mathbf{x}))^T$. Any function $g \in \mathcal{F}$ can then be written as a sum of these basis functions:

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{f}(\mathbf{x}) \,, \tag{2.5}$$

with some weight vector $\mathbf{w}$. The output signal $y(t) = g(\mathbf{x}(t))$ that the function produces from the training data can be written as a linear superposition of a nonlinearly expanded version $\mathbf{z}(t) = \mathbf{f}(\mathbf{x}(t))$ of the training data $\mathbf{x}$:

$$y(t) = \mathbf{w}^T \mathbf{f}(\mathbf{x}(t)) = \mathbf{w}^T \mathbf{z}(t) \,. \tag{2.6}$$

For simplicity, let us assume that the basis functions f_i are such that the expanded signals $\mathbf{z}$ have zero mean. If this is not the case, it can easily be achieved by subtracting the mean. In the above notation, the dependence of the Δ-value and the variance on the weight vector $\mathbf{w}$ can be made explicit

$$\Delta(y) := \langle \dot{y}(t)^2 \rangle_t \overset{(2.6)}{=} \mathbf{w}^T \underbrace{\langle \dot{\mathbf{z}}(t) \dot{\mathbf{z}}(t)^T \rangle_t}_{=:\dot{\mathbf{C}}} \mathbf{w} = \mathbf{w}^T \dot{\mathbf{C}} \mathbf{w} \,, \tag{2.7}$$

$$var(y) = \langle y(t)^2 \rangle_t = \mathbf{w}^T \underbrace{\langle \mathbf{z}(t) \mathbf{z}(t)^T \rangle_t}_{=:\mathbf{C}} \mathbf{w} = \mathbf{w}^T \mathbf{C} \mathbf{w} \,. \tag{2.8}$$

The first output signal to be found by SFA should minimize the Δ-value under the constraint of unit variance. The standard technique for such constrained optimization problems is the technique of Lagrange multipliers. The optima fulfill the necessary condition that the objective function

$$\mathcal{L}(\mathbf{w}) = \langle \dot{y}^2 \rangle_t - \lambda \langle y^2 \rangle_t = \mathbf{w}^T \dot{\mathbf{C}} \mathbf{w} - \lambda \mathbf{w}^T \mathbf{C} \mathbf{w} \tag{2.9}$$

should be stationary for some value of the Lagrange multiplier λ. By calculating the gradient of $\mathcal{L}$ and setting it to zero, we get the generalized eigenvalue problem

$$\dot{\mathbf{C}} \mathbf{w} = \lambda \mathbf{C} \mathbf{w} \,. \tag{2.10}$$

This generalized eigenvalue problem has a set of solutions $\mathbf{w}_j$ with eigenvalues λ_j. Let us assume that the eigenvectors $\mathbf{w}_j$ are sorted by ascending eigenvalue and that they are normalized according to $\mathbf{w}_j^T \mathbf{C} \mathbf{w}_j = 1$. It is straightforward to show that the functions $g_j(\mathbf{x}) = \mathbf{w}_j^T \mathbf{f}(\mathbf{x})$ are the solution to the optimization problem of SFA. Their Δ-value is given by the eigenvalue λ_j and they fulfill the constraints (2.2-2.4).

In summary, the solution to optimization problem 1 can be found by the following algorithm:

1. Calculate the nonlinear expansion $\mathbf{z}(t) = \mathbf{f}(\mathbf{x}(t))$ of the input data $\mathbf{x}(t)$ and assure that the expansion has zero mean.

2. Calculate the covariance matrix $\mathbf{C} = \langle \mathbf{z}\mathbf{z}^T \rangle_t$ of the expanded signals $\mathbf{z}(t)$ and the matrix $\dot{\mathbf{C}} = \langle \dot{\mathbf{z}}\dot{\mathbf{z}}^T \rangle_t$ of the second moments of its time derivative.

3. Solve the associated generalized eigenvalue problem (2.10).

The key advantage of this algorithm is that it is computationally efficient[1] and does not suffer from typical problems of gradient descent techniques such as local minima. The SFA algorithm has been implemented as part of a modular data processing toolbox (MDP, Berkes & Zito (2007)) in PYTHON and is readily available in the WWW.

[1] Linear algebra packages that efficiently solve generalized eigenvalue problems are readily available.

Part I

Slow Feature Analysis: Mathematical Approaches

Chapter 3

Finite-Dimensional Input Manifolds: Theory

3.1 Introduction

In this chapter, we will introduce a mathematical framework for the situation where the training data for SFA lie on a low-dimensional manifold, possibly embedded in a high-dimensional space. An example for such a situation is a video showing a single rotating object at a fixed position in space. No matter, how high-dimensional the pixel data may be, all pictures that can possibly occur in the video can be uniquely parameterized by three angles that characterize the orientation of the object in space. All pictures of the video thus lie on a 3-dimensional manifold embedded in the rather high-dimensional space of all possible images. Another example are videos that show optical input perceived by a virtual rat in a fixed virtual environment. The set of possible images can then be uniquely parameterized by the position of the rat and the direction of its gaze.

Of course, it is usually impossible to calculate the solutions of SFA for a particular video analytically. However, if we assume that SFA has access to an unrestricted function space $\mathcal{F}$, we can derive equations for the optimal functions and make analytical predictions for their behavior. In particular, the theory makes detailed predictions for the case where the input signals are generated from a set of statistically independent sources. In section 4.1, these results will be used to formulate a new algorithm for nonlinear blind source separation.

Although motivated by specific problems, the theory is formulated in a rather general fashion, so that it provides a framework that is hopefully also useful for other problems.

3.2 Basic Theory

In this section we will assume that the training data $\mathbf{x}$ for SFA are either finite-dimensional or lie on a finite-dimensional manifold, so that they can be parameterized by a finite-dimensional vector of parameters. In addition, we will assume that the function space $\mathcal{F}$ of SFA is unrestricted apart from sufficient differentiability and integrability. Although an unrestricted function space cannot be implemented in practice, it can serve as an abstraction of systems that implement complex function spaces, e.g. hierarchical systems (see section 4.2).

Representations of the Input Signals

The assumption that SFA has access to an unrestricted function space $\mathcal{F}$ has important theoretical implications. For restricted (but possibly still infinitely-dimensional) function spaces, coordinate changes in the space of the input data will in general alter the results, because they effectively change the function space from which the solutions are taken. As an example, assume that the input signal $\mathbf{x} = (x, y)$ is two-dimensional and the function space is the space of linear functions. Then, a change of the coordinate system to $(x', y') = (x^3, y)$ if still allowing only linear functions in the new coordinates leads to a very different function space. Thus the optimal functions will generate different optimal output signals y_j for the different coordinate systems. The optimization problem with a restricted function space is generally not invariant with respect to coordinate changes of the input signals.

For an unrestricted function space, the situation is different, because the concatenation of any function in $\mathcal{F}$ with the inversion of the coordinate change is again an element of the function space. The set of output signals that can be generated by the function space is then invariant with respect to coordinate changes of the input signals. Because the slowness of a function is measured in terms of its output signal, the optimal functions will of course depend on the coordinate system used, but the output signals will be the same.

This is particularly interesting in the situation described in the introduction, where the high-dimensional input signal does not cover the whole space of possible values, but lies on a low-dimensional manifold. For illustration, let us consider the example of the rotating object again. Because for unrestricted function spaces, the behavior of the optimal functions *outside* the input manifold is arbitrary, we are in general only interested in their behavior *on* the input manifold, that is, in the reaction of the system to all images that are possible within the given training scenario. The equivalence of different coordinate systems then implies that it is not important whether we take the (high-dimensional) video sequence or the (3-dimensional) time-dependent abstract angles as input signals. The output signal is the same. Of course the low-dimensional representation is much more amenable to analytical predictions and to intuitive interpretations of the system behavior. In section 4.2, we will use this simplification to predict the behavior of a hierarchical model of visual processing that reproduces the behavior of several cell types in the hippocampal formation of rodents commonly associated with spatial navigation (Franzius, Sprekeler & Wiskott, 2007).

Another situation in which the coordinate invariance is useful is the case of nonlinear blind source separation. Here, the input data are assumed to be a nonlinear mixture of some underlying sources. The task is to reconstruct the sources from the data without knowledge of the mixture or the sources. A natural prerequisite for the reconstruction is that the mixture is invertible. The mixture can then be interpreted as a nonlinear coordinate change, which is immaterial to the optimization problem above. From the theoretical perspective, we can thus simply assume that we had the sources as input signals and try to make predictions about how they are encoded in the optimal output signals found by SFA. If we can infer the sources (or good representatives thereof) from the optimal output signals under this condition, we can infer the sources from the output signals, no matter how they are encoded in the input data. Thus, SFA may be an interesting way of solving certain nonlinear blind source separation problems.

It is important to bear in mind that the theory developed in the following is valid

for an arbitrary choice of the input coordinate system, so that $\mathbf{x}(t)$ can stand for both concrete input signals (e.g. video sequences) or abstract representations of the input (e.g. angles that denote the orientation of the object in the video). Note however, that as the input data (or the manifold they lie on) becomes very high-dimensional, the resulting equations may be tedious to solve.

Further Assumptions and Notation

We assume that the input signal $\mathbf{x}(t)$ is ergodic, so that we can replace time averages by ensemble averages with a suitable probability density. Obviously, the optimization problem underlying SFA relies on the temporal structure of the training data as reflected by its derivative. Thus, a statistical description of the training signal $\mathbf{x}$ must incorporate not only the probability distribution for the values of $\mathbf{x}$, but rather the joint distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ of the input signal $\mathbf{x}$ and its derivative $\dot{\mathbf{x}}$.

We will assume that $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ is known and that we can define the marginal and conditional probability densities

$$p_{\mathbf{x}}(\mathbf{x}) := \int p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N\dot{x}\,, \tag{3.1}$$

$$p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x}) := \frac{p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})}{p_{\mathbf{x}}(\mathbf{x})}\,, \tag{3.2}$$

and the corresponding averages

$$\langle F(\mathbf{x},\dot{\mathbf{x}})\rangle_{\mathbf{x},\dot{\mathbf{x}}} := \int p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})F(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N x\,\mathrm{d}^N\dot{x}\,, \tag{3.3}$$

$$\langle F(\mathbf{x})\rangle_{\mathbf{x}} := \int p_{\mathbf{x}}(\mathbf{x})F(\mathbf{x})\,\mathrm{d}^N x\,, \tag{3.4}$$

$$\langle F(\mathbf{x},\dot{\mathbf{x}})\rangle_{\dot{\mathbf{x}}|\mathbf{x}}(\mathbf{x}) := \int p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x})F(\mathbf{x},\dot{\mathbf{x}})\,\mathrm{d}^N\dot{x}\,. \tag{3.5}$$

Throughout the thesis, we will assume that all averages taken exist. This introduces integrability constraints on the functions of which the average is taken. The function space is thus not completely unrestricted. The functions are restricted to be integrable in the sense that the averages above exist. In addition, they should be differentiable, simply to assure that the temporal derivative of their output signal exists.

Partial derivatives with respect to x_μ will be written as ∂_μ. For example, the divergence of a vector field $\mathbf{v}(\mathbf{x})$ then takes the short form

$$\operatorname{div}\mathbf{v}(\mathbf{x}) := \sum_\mu \frac{\partial v_\mu(\mathbf{x})}{\partial x_\mu} = \sum_\mu \partial_\mu v_\mu(\mathbf{x})\,. \tag{3.6}$$

We use the convention that within products, ∂_μ acts on all functions to its right. If we want ∂_μ to act locally, we use angular brackets. This convention can be illustrated by the product rule

$$\partial_\mu F(\mathbf{x})G(\mathbf{x}) = [\partial_\mu F(\mathbf{x})]G(\mathbf{x}) + F(\mathbf{x})[\partial_\mu G(\mathbf{x})]\,. \tag{3.7}$$

Reformulation of the Optimization Problem

To describe the Δ-value in terms of the probability distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$, we need to express the temporal derivative of the output signal $y(t) = g(\mathbf{x}(t))$ in terms of the input signals $\mathbf{x}$ and their derivatives. This is readily done by the chain rule

$$\dot{y}(t) = \frac{\mathrm{d}}{\mathrm{d}t} g(\mathbf{x}(t)) = \sum_{\mu} \dot{x}_\mu(t) \partial_\mu g(\mathbf{x}(t)) \,. \tag{3.8}$$

We can now rewrite the objective function (2.1) by replacing the time average $\langle\cdot\rangle_t$ by the ensemble average $\langle\cdot\rangle_{\mathbf{x},\dot{\mathbf{x}}}$

$$\begin{aligned}
\Delta(g_j) &\overset{(2.1)}{=} \langle \dot{y}(t)^2 \rangle_t && (3.9)\\
&\overset{(3.8)}{=} \sum_{\mu,\nu} \langle \dot{x}_\mu [\partial_\mu g_j(\mathbf{x})] \dot{x}_\nu [\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x},\dot{\mathbf{x}}} && (3.10)\\
&= \sum_{\mu,\nu} \langle \underbrace{\langle \dot{x}_\mu \dot{x}_\nu \rangle_{\dot{\mathbf{x}}|\mathbf{x}}}_{=:K_{\mu\nu}(\mathbf{x})} [\partial_\mu g_j(\mathbf{x})][\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x}} && (3.11)\\
&= \sum_{\mu,\nu} \langle K_{\mu\nu}(\mathbf{x}) [\partial_\mu g_j(\mathbf{x})][\partial_\nu g_j(\mathbf{x})] \rangle_{\mathbf{x}} \,, && (3.12)
\end{aligned}$$

where $K_{\mu\nu}(\mathbf{x})$ is the matrix of the second moments of the conditional velocity distribution $p_{\dot{\mathbf{x}}|\mathbf{x}}(\dot{\mathbf{x}}|\mathbf{x})$ and reflects the dynamical structure of the input signal.

An elegant reformulation of the optimization problem can be obtained by introducing the following scalar product (f,g) between functions $f, g \in \mathcal{F}$:

$$(f,g) = \langle f(\mathbf{x}) g(\mathbf{x}) \rangle_{\mathbf{x}} \,. \tag{3.13}$$

With this definition, the function space $\mathcal{F}$ becomes a Hilbert space and the slowness objective for a function g can be written as

$$\Delta(g) = \sum_{\mu,\nu} (\partial_\mu g, K_{\mu\nu} \partial_\nu g) \,. \tag{3.14}$$

Note that we restrict the action of the partial derivatives to the argument of the scalar product they appear in.

Replacing the temporal averages by ensemble averages and using the scalar product (3.13), the original optimization problem becomes

Optimization problem 2: *Given a function space $\mathcal{F}$ and a probability distribution $p_{\mathbf{x},\dot{\mathbf{x}}}(\mathbf{x},\dot{\mathbf{x}})$ for the input signal $\mathbf{x}$ and its derivative $\dot{\mathbf{x}}$, find a set of $J+1$ real-valued input-output functions $g_j(\mathbf{x}), j \in \{0,1,...,J\}$ that minimize*

$$\Delta(g_j) = \sum_{\mu,\nu} (\partial_\mu g_j, K_{\mu\nu} \partial_\nu g_j) \tag{3.15}$$

under the constraint

$$\forall i < j: \quad (g_i, g_j) = \delta_{ij} \qquad \textit{(orthonormality)} \,. \tag{3.16}$$

Here we dropped the zero mean constraint and allow the trivial constant solution $g_0 = 1$ to occur. As any function whose scalar product with the constant vanishes must have zero mean, the constraint (3.16) implies zero mean for all functions with $j > 0$. For functions f, g with zero mean, in turn, the scalar product (3.13) is simply the covariance, so that the constraints (2.2-2.4) can be compactly written as the orthonormality constraint (3.16).

A Differential Equation for the Solutions

In this section we will show that optimization problem 2 can be reduced to a partial differential eigenvalue equation. As some of the proofs are lengthy and not very illustrative, we will state and motivate the main results while postponing the exact proofs to the appendix.

Under the assumption that all functions $g \in \mathcal{F}$ fulfill a boundary condition that will be stated below, the objective function can be written as

$$\Delta(g) = (g, \underbrace{\sum_{\mu,\nu} \partial_\mu^\dagger K_{\mu\nu} \partial_\nu}_{=:\mathcal{D}} g) = (g, \mathcal{D}g)\,. \tag{3.17}$$

Here, $A^\dagger$ denotes the adjoint operator to A with respect to the scalar product (3.13), i.e., the operator that fulfills the condition $(Af, g) = (f, A^\dagger g)$ for all functions $f, g \in \mathcal{F}$. It can be shown that $\partial_\mu^\dagger = -\frac{1}{p_\mathbf{x}} \partial_\mu p_\mathbf{x}$. Thus the operator $\mathcal{D}$ is the partial differential operator

$$\mathcal{D} = -\frac{1}{p_\mathbf{x}} \sum_{\mu,\nu} \partial_\mu p_\mathbf{x} K_{\mu\nu} \partial_\nu\,. \tag{3.18}$$

Because $K_{\mu\nu}$ is symmetric, $\mathcal{D}$ is self-adjoint, i.e. $(f, \mathcal{D}g) = (\mathcal{D}f, g)$.

The main advantage of this reformulation is that the Δ-value takes a form that is common in other contexts, e.g. in quantum mechanics, where the operator $\mathcal{D}$ corresponds to the Hamilton operator (e.g. Landau & Lifshitz, 1977, §20). This analogy allows us to transfer the well-developed theory from these areas to our problem. As in quantum mechanics, the central role is played by the eigenfunctions of $\mathcal{D}$. This culminates in theorem 1, which we will briefly motivate. A rigorous proof can be found in appendix A and in the supplementary to Franzius, Sprekeler & Wiskott (2007).

Because the operator $\mathcal{D}$ is self-adjoint, it possesses a complete set of eigenfunctions g_i that are mutually orthogonal with respect to the scalar product (3.13) (spectral theorem, see e.g. Courant & Hilbert, 1989, chapter V, §14). The eigenfunctions of $\mathcal{D}$ are defined by the eigenvalue equation

$$\mathcal{D}g_i = \lambda_i g_i \tag{3.19}$$

and are assumed to be normalized according to

$$(g_i, g_i) = 1\,. \tag{3.20}$$

Because they are orthogonal, they thus fulfill the orthonormality constraint (3.16). Inserting these expressions into (3.17) immediately shows that the Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_i) \overset{(3.17)}{=} (g_i, \mathcal{D}g_i) \overset{(3.19)}{=} (g_i, \lambda_i g_i) \overset{(3.20)}{=} \lambda_i\,. \tag{3.21}$$

Because of the completeness of the eigenfunctions g_i, any function g can be represented as a linear combination $g = \sum_i w_i g_i$ of the eigenfunctions g_i. The Δ-value of g can then be decomposed into a sum of the Δ-values of the eigenfunctions

$$\Delta(g) = (g, \mathcal{D}g) \overset{(3.21,3.16)}{=} \sum_i w_i^2 \lambda_i \,. \tag{3.22}$$

The unit variance constraint requires that the square sum of the coefficients w_i is unity: $\sum_i w_i^2 = 1$. It is then evident that the Δ-value (3.22) can be minimized by choosing $w_i = \delta_{0i}$, so that the slowest function is simply the eigenfunction g_0 with the smallest eigenvalue. The space of all functions that are orthogonal to g_0 is spanned by the remaining eigenfunctions g_i with $i > 0$. The slowest function in this space is the eigenfunction g_1 with the second smallest eigenvalue. Iterating this scheme makes clear that the optimal functions are simply the eigenfunctions g_i, ordered by their eigenvalue.

A detailed analysis of the problem shows that the optimal functions are given by the eigenfunctions that fulfill von Neumann boundary conditions (see Appendix A).

Theorem 1. *The solution of optimization problem 2 is given by the J eigenfunctions of the operator $\mathcal{D}$ with the smallest eigenvalues, i.e. the functions that fulfill*

$$\mathcal{D}g_i = \lambda_i g_i \tag{3.23}$$

with the boundary condition

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_i = 0 \tag{3.24}$$

and the normalization condition

$$(g_i, g_i) = 1 \,. \tag{3.25}$$

Here, $\mathbf{n}(\mathbf{x})$ is the normal vector on the boundary for the point $\mathbf{x}$. The Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_i) = \lambda_i \,. \tag{3.26}$$

If the input data $\mathbf{x}$ are not bounded, the boundary condition has to be replaced by a limit, so that for parameterized boundaries that grow to infinity, the left hand side of equation (3.24) converges to zero for all points on the boundary. Note that we assumed earlier that all averages taken exist. This implies that the square of the functions and their first derivatives decay more quickly than $p_{\mathbf{x}}(\mathbf{x})$ as $||\mathbf{x}|| \to \infty$. Functions that do not fulfill the limit case of the boundary condition tend to have infinite variance or Δ-value.

The key advantage of the developed theory is that it converts the (global) optimization problem into a (local) partial differential eigenvalue equation. Moreover, the eigenvalue equation (3.23) belongs to a class that is known as Sturm-Liouville problems (see e.g. Courant & Hilbert (1989)), for which a well-developed theory exists. In the next chapter we will use Sturm-Liouville theory to study the case of input data generated from a set of statistically independent sources.

3.3 Statistically Independent Sources

Motivation: Nonlinear Blind Source Separation

The task of blind source separation (BSS) is to extract a set of S underlying sources $\mathbf{s} \in \mathbb{R}^S$ from a given set of data $\mathbf{x} \in \mathbb{R}^N$ that were generated from the sources by means of some unknown invertible mixture $\mathbf{x} = \mathbf{F}(\mathbf{s})$. Of course this task cannot be solved without additional knowledge of the properties of the sources. Typically, one assumption is that they are statistically independent. In the case of a linear relation between the sources and the data (i.e., $\mathbf{x} = \mathbf{F}(\mathbf{s}) = \mathbf{A}\mathbf{s}$ with some matrix $\mathbf{A}$) and if at most one of the sources is Gaussian, statistical independence is a sufficient criterion for the recovery of the sources (up to scaling and permutation). Several techniques for BSS rely on this assumption (see Hyvärinen et al., 2001, for an overview).

Unfortunately, in the case of general nonlinear mixtures, statistical independence of the sources is not a sufficient criterion for the problem. Because any nonlinear transformation of a single source is still statistically independent of the other sources, there is an infinite number of functions that succeed in generating statistically independent signals but fail to recover the underlying sources. Moreover, it has been shown that it is even possible to construct nonlinear mixtures of the sources that lead to statistically independent output signals (Hyvärinen & Pajunen, 1999). Additional constraints have to be imposed to resolve these ambiguities.

In this section, we use the theory developed in the previous section to make predictions on how an SFA system with an unrestricted function space should behave in the case where the input data are a nonlinear mixture of a set of statistically independent sources. As discussed in section 3.2, the nonlinear mixture can be regarded as a coordinate change of the input data and is thus immaterial to the output signals generated by SFA. Consequently, it makes no difference if we use the nonlinear mixture as the input signals or the undistorted and unmixed sources. The output signals should be the same. Interestingly, the statistical independence of the sources leads to a particular representation of the sources in the output signals of SFA.

Factorization of the Output Signals

In the following we assume that the input signals for SFA are the sources $\mathbf{s} \in \mathbb{R}^S$. To emphasize that the input signals are the sources and not the mixture, we will use indices α and β for the sources instead of μ and ν for the components of the input signals. The statistical independence of the sources is formally reflected by the factorization of their joint probability density

$$p_{\mathbf{s},\dot{\mathbf{s}}}(\mathbf{s},\dot{\mathbf{s}}) = \prod_{\alpha} p_{s_\alpha,\dot{s}_\alpha}(s_\alpha,\dot{s}_\alpha)\,. \tag{3.27}$$

Then, the marginal probability $p_{\mathbf{s}}$ also factorizes into the individual probabilities $p_\alpha(s_\alpha)$

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_{\alpha} p_\alpha(s_\alpha) \tag{3.28}$$

and $K_{\alpha\beta}$ is diagonal

$$K_{\alpha\beta}(\mathbf{s}) = \delta_{\alpha\beta} K_\alpha(s_\alpha) \quad \text{with} \quad K_\alpha(s_\alpha) := \langle \dot{s}_\alpha^2 \rangle_{\dot{s}_\alpha|s_\alpha}\,. \tag{3.29}$$

The latter is true because the mean temporal derivative of 1-dimensional stationary and differentiable stochastic processes vanish for any s_α for continuity reasons, so that $K_{\alpha\beta}$ is

not only the matrix of the second moments of the derivatives, but actually the conditional covariance matrix of the derivatives of the sources given the sources. As the sources are statistically independent, their derivatives are uncorrelated and $K_{\alpha\beta}$ has to be diagonal.

The operator $\mathcal{D}$ introduced in section 3.2 can then be split into a sum of operators $\mathcal{D}_\alpha$, each of which depends on only one of the sources:

$$\mathcal{D}(\mathbf{s}) = \sum_\alpha \mathcal{D}_\alpha(s_\alpha) \tag{3.30}$$

with

$$\mathcal{D}_\alpha = -\frac{1}{p_\alpha}\partial_\alpha p_\alpha K_\alpha \partial_\alpha \,. \tag{3.31}$$

This has the important implication that the solution to the full eigenvalue problem for $\mathcal{D}$ can be constructed from the 1-dimensional eigenvalue problems associated with $\mathcal{D}_\alpha$:

Theorem 2. *Let $g_{\alpha i}$ ($i \in \mathbb{N}$) be the normalized eigenfunctions of the operators $\mathcal{D}_\alpha$, i.e., the set of functions $g_{\alpha i}$ that fulfill the eigenvalue equations*

$$\mathcal{D}_\alpha g_{\alpha i} = \lambda_{\alpha i} g_{\alpha i} \tag{3.32}$$

with the boundary conditions

$$p_\alpha K_\alpha \partial_\alpha g_{\alpha i} = 0 \tag{3.33}$$

and the normalization condition

$$(g_{\alpha i}, g_{\alpha i})_\alpha := \langle g_{\alpha i}^2 \rangle_\alpha = 1 \,. \tag{3.34}$$

Then, the product functions

$$g_{\mathbf{i}}(\mathbf{s}) := \prod_\alpha g_{\alpha i_\alpha}(s_\alpha) \tag{3.35}$$

form a complete set of (normalized) eigenfunctions to the full operator $\mathcal{D}$ with the eigenvalues

$$\lambda_{\mathbf{i}} = \sum_\alpha \lambda_{\alpha i_\alpha} \tag{3.36}$$

and thus those $g_{\mathbf{i}}$ with the smallest eigenvalues $\lambda_{\mathbf{i}}$ form a solution of optimization problem 2. Here, $\mathbf{i} = (i_1, ..., i_S) \in \mathbb{N}^S$ denotes a multi-index that enumerates the eigenfunctions of the full eigenvalue problem.

In the following, we will assume that the eigenfunctions $g_{\alpha i}$ are ordered by their eigenvalue and refer to them as the *harmonics* of the source s_α. This is motivated by the observation that in the case where p_α and K_α are independent of s_α, i.e., for a uniform distribution, the eigenfunctions $g_{\alpha i}$ are harmonic oscillations whose frequency increases linearly with i (for a derivation see below). Moreover, we will assume that the sources s_α are ordered according to slowness, in this case measured by the eigenvalue $\lambda_{\alpha 1}$ of their lowest non-constant harmonic $g_{\alpha 1}$. These eigenvalues are the Δ-value of the slowest possible nonlinear point transformations of the sources.

The main result of the above theorem is that in the case of statistically independent sources, the output signals are products of harmonics of the sources. Note that the constant function $g_{\alpha 0}(s_\alpha) = 1$ is an eigenfunction with eigenvalue 0 to all the eigenvalue problems (3.32). As a consequence, the harmonics $g_{\alpha i}$ of the single sources are also eigenfunctions to the full operator $\mathcal{D}$ (with the index $\mathbf{i} = (0, ..., 0, i_\alpha = i, 0, ..., 0)$) and

can thus be found by SFA. Importantly, the lowest non-constant harmonic of the slowest source (i.e., $g_{(1,0,0,\ldots)} = g_{11}$) is the function with the smallest overall Δ-value (apart from the constant) and thus the first function found by SFA. In the next sections, we will show that the lowest non-constant harmonics reconstruct the sources up to a monotonic and thus invertible point transformation and that in the case of sources with Gaussian statistics, they even reproduce the sources exactly.

Monotony of the First Harmonic

Let us assume that the source s_α is bounded and takes on values on the interval $s_\alpha \in [a, b]$. The eigenvalue problem (3.32,3.33) can be rewritten in the standard form of a Sturm-Liouville problem

$$\partial_\alpha p_\alpha K_\alpha \, \partial_\alpha g_{\alpha i} + \lambda_{\alpha i} p_\alpha g_{\alpha i} = 0 \,, \tag{3.37}$$

$$p_\alpha K_\alpha \partial_\alpha g_{\alpha i} = 0 \quad \text{for} \quad s_\alpha \in \{a, b\} \,. \tag{3.38}$$

Note that both p_α and $p_\alpha K_\alpha$ are positive for all s_α. Sturm-Liouville theory states that the solutions $g_{\alpha i}, i \in \mathbb{N}^0$ of this problem are oscillatory and that $g_{\alpha i}$ has exactly i zeros on $]a, b[$, if the $g_{\alpha i}$ are ordered by increasing eigenvalue $\lambda_{\alpha i}$ (Courant & Hilbert, 1989, chapter IV, §6). All eigenvalues are positive. In particular, $g_{\alpha 1}$ has only one zero $\xi \in]a, b[$. Without loss of generality we assume that $g_{\alpha 1} < 0$ for $s_\alpha < \xi$ and $g_{\alpha 1} > 0$ for $s_\alpha > \xi$. Then equation (3.37) implies that

$$\partial_\alpha p_\alpha K_\alpha \partial_\alpha g_{\alpha 1} = -\lambda_\alpha p_\alpha g_{\alpha 1} < 0 \quad \text{for } s_\alpha > \xi \tag{3.39}$$

$$\Rightarrow p_\alpha K_\alpha \partial_\alpha g_{\alpha 1} \quad \text{is monotonic decreasing on }]\xi, b] \tag{3.40}$$

$$\overset{(3.38)}{\Rightarrow} p_\alpha K_\alpha \, \partial_\alpha g_{\alpha 1} > 0 \text{ on }]\xi, b[\tag{3.41}$$

$$\overset{p_\alpha K_\alpha > 0}{\Rightarrow} \partial_\alpha g_{\alpha 1} > 0 \text{ on }]\xi, b[\tag{3.42}$$

$$\Leftrightarrow g_{\alpha 1} \quad \text{is monotonic increasing on }]\xi, b[\,. \tag{3.43}$$

A similar consideration for $s < \xi$ shows that $g_{\alpha 1}$ is also monotonic increasing on $]a, \xi[$. Thus, $g_{\alpha 1}$ is monotonic and invertible on the whole interval $[a, b]$. Note that the monotony of $g_{\alpha 1}$ is important in the context of BSS, because it ensures that not only some of the output signals of SFA depend on only one of the sources (the harmonics), but that there should actually be some that are very similar to the source itself (the lowest non-constant harmonics).

Gaussian Sources

We will now consider the situation that the sources are stationary Gaussian stochastic processes. Because of the stationarity, the sources and their temporal derivatives are statistically independent, i.e., $p_{\dot{s}_\alpha | s_\alpha}(\dot{s}_\alpha | s_\alpha) = p_{\dot{s}_\alpha}(\dot{s}_\alpha)$. Thus, K_α is independent of s_α, i.e., $K_\alpha(s_\alpha) = K_\alpha = const.$ Without loss of generality we assume that the sources have unit variance. Then the probability density of the source is given by

$$p_\alpha(s_\alpha) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s_\alpha^2}{2}\right) \tag{3.44}$$

and the eigenvalue equations (3.37) for the harmonics can be written as

$$\partial_\alpha e^{-s_\alpha^2/2}\partial_\alpha g_{\alpha i} + \frac{\lambda_{\alpha i}}{K_\alpha} e^{-s_\alpha^2/2} g_{\alpha i} = 0\,. \tag{3.45}$$

This is a standard form of Hermite's differential equation (see Courant & Hilbert, 1989, chapter V, §10). Accordingly, the harmonics $g_{\alpha i}$ are given by the (appropriately normalized) Hermite polynomials H_i of the sources:

$$g_{\alpha i}(s_\alpha) = \frac{1}{\sqrt{2^i i!}} H_i\left(\frac{s_\alpha}{\sqrt{2}}\right)\,. \tag{3.46}$$

The Hermite polynomials can be expressed in terms of derivatives of the Gaussian distribution:

$$H_n(x) = (-1)^n \mathrm{e}^{x^2} \partial_x^n \mathrm{e}^{-x^2}\,. \tag{3.47}$$

It is clear that Hermite polynomials fulfill the boundary condition

$$\lim_{s_\alpha \to \infty} K_\alpha p_\alpha \partial_\alpha g_{\alpha i} = 0\,, \tag{3.48}$$

because the derivative of a polynomial is again a polynomial and a Gaussian distribution decays faster than polynomially as $|s_\alpha| \to \infty$. The eigenvalues are given by

$$\lambda_{\alpha i_\alpha} = i_\alpha K_\alpha\,. \tag{3.49}$$

The most important consequence is that the lowest non-constant harmonics simply reproduce the sources: $g_{\alpha 1}(s_\alpha) = 1/\sqrt{2} H_1(s_\alpha/\sqrt{2}) = s_\alpha$. Thus, for Gaussian sources, SFA with an unrestricted function space will reproduce the sources, although it still remains to determine which of the output signals are the sources and which are higher harmonics or products of the harmonics of the sources. In section 4.1 we will present an algorithm that should ideally solve this problem.

Homogeneously Distributed Sources

Another canonical example for which the eigenvalue equation (3.32) can be solved analytically is the case of homogeneously distributed sources, i.e. the case where the probability distribution $p_{s,\dot{s}}$ is independent of s. Consequently, neither $p_\alpha(s_\alpha)$ nor $K_\alpha(s_\alpha)$ can depend on s_α, i.e., they are constants. Note that such a distribution may be difficult to implement by a real differentiable process, because the velocity distribution should be different at boundaries that cannot be crossed. Nevertheless, this case provides an approximation to cases, where the distribution is close to homogeneous.

Let s_α take values in the interval $[0, L_\alpha]$. The eigenvalue equation for the harmonics is then given by

$$-K_\alpha \partial_\alpha^2 g_{\alpha i} = \lambda_{\alpha i} g_{\alpha i}\,. \tag{3.50}$$

and readily solved by harmonic oscillations:

$$g_{\alpha i}(s_\alpha) = \sqrt{2}\cos\left(i\pi \frac{s_\alpha}{L_\alpha}\right)\,. \tag{3.51}$$

The Δ-value of these functions is given by

$$\Delta(g_{\alpha i}) = \lambda_{\alpha i} = K_\alpha \left(\frac{\pi}{L_\alpha} i\right)^2\,. \tag{3.52}$$

Weakly Inhomogeneous Sources

For homogeneous distributions, the optimal functions for SFA are harmonic oscillations. It is reasonable to assume that this behavior will be preserved qualitatively if p_α and K_α are no longer homogeneous but depend weakly on the source s_α. In particular, if the wavelength of the oscillation is much shorter than the typical scale on which p_α and K_α vary, it can be expected that the oscillation "does not notice" the change. Of course, we are not principally interested in quickly varying functions, but they can provide insights into the effect of variations of p_α and K_α.

To examine this further, we can derive an approximate solution to the eigenvalue equation (3.32,3.37) by treating $\epsilon = 1/\sqrt{\lambda_{\alpha i}} = 1/\sqrt{\Delta}$ as a small perturbation parameter. This corresponds to large Δ-values, i.e., quickly varying functions. For this case we can apply a perturbation theoretical approach that follows the Wentzel-Kramers-Brillouin approximation used in quantum mechanics. For a more detailed description of the approach we refer the reader to the quantum mechanical literature (e.g. Davydov, 1976). Knowing that the solution shows oscillations, we start with the complex ansatz

$$g_\alpha(s_\alpha) = A \exp\left(\frac{i}{\epsilon}\Phi(s_\alpha)\right) , \tag{3.53}$$

where $\Phi(s_\alpha)$ is a complex function that needs to be determined. Treating ϵ as a small number, we can expand Φ in order of ϵ

$$\Phi(s_\alpha) = \Phi_0(s_\alpha) + \epsilon\Phi_1(s_\alpha) + \ldots , \tag{3.54}$$

where the ellipses stand for higher-order terms. We insert this expression into the eigenvalue equation (3.37) and collect terms of the same order in ϵ. Requiring each order to vanish separately and neglecting orders of ϵ^2 and higher, we get equations for Φ_0 and Φ_1:

$$(\partial_\alpha\Phi_0)^2 = \frac{1}{K_\alpha} , \tag{3.55}$$

$$\partial_\alpha\Phi_1 = \frac{i}{2}\frac{\partial_\alpha(p_\alpha K_\alpha \partial_\alpha\Phi_0)}{p_\alpha K_\alpha \partial_\alpha\Phi_0} . \tag{3.56}$$

These equations are solved by

$$\Phi_0(s_\alpha) = \int_{s_0}^{s_\alpha} \sqrt{\frac{1}{K_\alpha(s)}}\,\mathrm{d}s , \tag{3.57}$$

$$\Phi_1(s_\alpha) = \frac{i}{2}\ln\left(p_\alpha K_\alpha^{1/2}\right) , \tag{3.58}$$

where s_0 is an arbitrary reference point. Inserting this back into equation (3.53), we get the approximate solution

$$g_\alpha(s_\alpha) = \frac{A}{\sqrt[4]{p_\alpha^2 K_\alpha}} \exp\left(i\int_{s_0}^{s_\alpha} \sqrt{\frac{\Delta}{K_\alpha(s)}}\mathrm{d}s\right) . \tag{3.59}$$

This shows that the solutions with large Δ-values show oscillations with local frequency $\sqrt{\Delta/K_\alpha}$ and amplitude $\sim 1/\sqrt[4]{p_\alpha^2 K_\alpha}$. Large values of K_α indicate that the source changes quickly, i.e., where its "velocity" is high. This implies that the local frequency of the solutions is smaller for values of the sources where the source velocity is high, whereas

small source velocities lead to higher frequencies than expected for homogeneous movement. Intuitively, this means that the functions compensate for high source velocities with smaller spatial frequencies such that the effective temporal frequency of the output signal is kept constant.

Understanding the dependence of the amplitude on p_α and K_α is more subtle. Under the assumption that K_α is independent of s_α, the amplitude decreases where p_α is large and increases where p_α is small. Intuitively, this can be interpreted as an equalization of the fraction of the total variance that falls into a small interval of length $\Delta s_\alpha \gg \sqrt{K_\alpha/\Delta}$. This fraction is roughly given by the product of the probability $p_\alpha \Delta s_\alpha$ of being in this section times the squared amplitude $1/\sqrt{p_\alpha^2 K_\alpha}$ of the oscillation. For constant K_α, this fraction is also constant, so the amplitude is effectively rescaled to yield the same "local variance" everywhere. If p_α is constant and K_α varies, on the other hand, the amplitude of the oscillation is small for values of the sources where they change quickly and large where they change slowly. This corresponds to the intuition that there are two ways of treating regions where the sources change quickly: decreasing spatial frequency to generate slower output signals and/or decreasing the amplitude of the oscillation to "pay less attention" to these regions. There is also a strong formal argument why the amplitude should depend on $p_\alpha^2 K_\alpha$. As the optimization problem is invariant under arbitrary invertible coordinate changes, the amplitude of the oscillation should depend on a function of p_α and K_α that is independent of the coordinate system. This constrains the amplitude to depend on $p_\alpha^2 K_\alpha$, as this is the only combination of these quantities that is invariant under coordinate changes.

In simulations, we expect that the effects of inhomogeneities in the distribution will be reflected mainly by variations in the spatial frequency of the solutions. For continuous trajectories of the sources, the probability for a source to take on values in an interval $[s_\alpha, s_\alpha + \mathrm{d}s]$ should be proportional to the time the source remains within this interval. This time, in turn, should be inversely proportional to the velocity of the signal. Intuitively, one could thus expect that the variance K_α of the velocity is inversely proportional to p_α^2, which would yield $p_\alpha^2 K_\alpha$ and thus the amplitude of the oscillation constant.

3.4 Analogies in Physics

Slow Feature Analysis and Hamilton's Principle

The last two sections as well as previous studies (Wiskott, 2003) have illustrated that SFA allows a rich repertoire of analytical considerations. Why is that? The main reason is that both the Δ-value and the constraints are quadratic functionals of the output signals. As long as the output signal is linearly related to the parameters of the input-output functions (as is the case for the nonlinear expansion approach that underlies the SFA algorithm), both the Δ-value and the constraint quantities are quadratic forms of the parameters. The gradients involved in finding the optima are thus linear functions of the parameters, so that the solution can be found by means of linear methods, typically eigenvalue problems.

Eigenvalue problems have a long history in mathematical physics. They describe electron orbitals in atoms, acoustic resonances, vibrational modes in solids and light propagation in optical fibers. Whenever wave phenomena are involved, the associated theory makes use of eigenvalue problems in one way or another. Consequently, there

is a well-developed mathematical theory for eigenvalue problems, including the infinite-dimensional case.

SFA aims at minimizing the mean square of the temporal derivative of the output signal y. Let us assume for a moment that we were only interested in the first, i.e., the slowest output signal of SFA. Then the only constraint that applies is that of unit variance. According to the technique of Lagrange multipliers, we are searching for stationary points of the objective function

$$\mathcal{L}(y) = \langle \dot{y}^2 \rangle_t - \lambda \langle y^2 \rangle_t = \frac{1}{T} \int \dot{y}(t)^2 \mathrm{d}t - \frac{\lambda}{T} \int y(t)^2 \mathrm{d}t \,, \tag{3.60}$$

where λ is a Lagrange multiplier, which has to be determined such that the constraints are fulfilled.

To interpret this objective function let us for a moment act as if the output signal y was the position of a physical particle. Then the square of the temporal derivative of y is proportional to the kinetic energy of the particle. We can thus interpret $K = \dot{y}^2/T$ as the *kinetic energy* of the output signal y. Consequently, it is only natural to interpret the second term in equation (3.60) in terms of a *potential energy* $U = \lambda y^2/T$. Then, the objective function $\mathcal{L}$ is the integral over the difference between the kinetic and the potential energy of the output signal, a quantity that is known as the *action* in Lagrange mechanics:

$$\mathcal{L}(y) = \int \left[K(t) - U(t) \right] \mathrm{d}t \,. \tag{3.61}$$

One of the most important principles of Lagrange mechanics is Hamilton's principle of least action, which states that out of all possible trajectories, physical systems "choose" those for which the action $\mathcal{L}$ is stationary. It is immediately clear that with the above reinterpretation of the quantities appearing in SFA, the two problems are formally very similar.

Moreover, since the potential energy of the physical system corresponding to SFA is quadratic in the "position y of the particle", the problem is in essence that of a harmonic oscillator. From this perspective, it is not surprising that the optimal output signals for SFA are generally harmonic oscillations, as shown by Wiskott (2003).

In these considerations, we have neglected that the output signals cannot take an arbitrary shape, because they are determined by the input data and some input-output function g. We will see below, however, that formal analogies remain even when this is taken into account.

Standing Waves

The optimal solutions for SFA are given by the eigenfunctions of the operator $\mathcal{D}$, which is a quadratic form in the partial derivatives ∂_μ. Hence, $\mathcal{D}$ belongs to the same class of operators as the Laplace operator. This implies that equation (3.23) has the form of a stationary wave equation, which describes oscillatory eigenmodes of fields.

An intuitive picture can be sketched for the exemplary case that the input data $\mathbf{x}$ lies on a 2-dimensional manifold, embedded in a 3-dimensional space. We can then interpret this manifold as an oscillating membrane. Equation (3.23) describes the vibrational eigenmodes or standing waves on the membrane. The boundary condition (3.24) means that the boundary of the membrane is open, i.e. the borders of the membrane can oscillate freely. The solutions $g_i(\mathbf{x})$ of SFA correspond to the amplitude of an eigenmode

with frequency $\omega = \sqrt{\lambda_i}$ at position $\mathbf{x}$. For a constant probability distribution $p_{\mathbf{x}}$, the matrix $K_{\mu\nu}$ can moreover be interpreted as the "surface tension" of the membrane. In a given direction $\mathbf{n}$ ($\mathbf{n}$ tangential to the membrane and $|\mathbf{n}| = 1$), the "tension" of the membrane is given by $\kappa = n_\mu K_{\mu\nu} n_\nu$. If the input changes quickly in the direction of $\mathbf{n}$, the surface tension κ is large. For large surface tension, however, oscillations with a given wavelength have a high frequency, that is, a large Δ-value. Thus, slow functions (solutions with small Δ-values $\sim$ oscillations with low frequency) will tend be oscillatory in directions with small input velocity (low surface tension) and remain largely constant in directions of large input velocity (high surface tension). Directions with high surface tension correspond to input directions in which SFA will learn invariances.

Quantum Mechanics

An intuition for the factorization of the solutions for independent sources can be gained by interpreting $\mathcal{D}$ as a formal equivalent of the Hamilton operator in quantum mechanics. Equation (3.23) then corresponds to the stationary Schrödinger equation and the Δ-values λ_i to the energies of stationary states of a quantum system. For statistically independent sources, the operator $\mathcal{D}$ decomposes into a sum of operators $\mathcal{D}_\mu$, which depend on only one of the sources each. The decomposition corresponds to the situation of a quantum system with "Hamilton operator" $\mathcal{D}$ that consists of a set of independent quantum systems with "Hamilton operators" $\mathcal{D}_\mu$. For readers who are familiar with quantum mechanics, it is then no longer surprising that the eigenvalue equation for $\mathcal{D}$ can be solved by means of a separation ansatz. The solutions of SFA (stationary states of the full quantum system) are thus products of the harmonics of the sources in isolation (the stationary states of the independent subsystems). Similarly, it is clear that the Δ-value of the product states (the energy of the full system) is the sum of the Δ-values of the harmonics (the energies of the subsystems).

The dependence of the Δ-value (energy) $\lambda_{\mathbf{i}}$ on the index (quantum number) $\mathbf{i}$ also has a counterpart in physics. As a function of the source s_μ, the harmonics $g_{\mu i_\mu}$ show oscillations with i_μ zeros. Thus, the index i_μ is a measure of the spatial frequency (or, in quantum mechanics, the momentum) of the harmonic. From this perspective, the dependence of the Δ-value (energy) on the index (frequency or momentum) $\mathbf{i}$ plays the role of a dispersion relation. For homogeneously distributed sources, the dispersion is quadratic, while for Gaussian sources it is linear.

Wave equations of the type of equation (3.23) are ubiquitous in physics and there are probably more formally equivalent physical systems. We believe that these analogies can help substantially in getting an intuitive understanding of the behavior of SFA in the limit case of very rich function spaces.

Chapter 4

Finite-Dimensional Input Manifolds: Applications

In this chapter we will present two applications of the theory developed in chapter 3. First, in section 4.1, we will present a novel algorithm for nonlinear blind source separation and test its performance for mixtures of audio data. Then, in section 4.2, we will analyze the behavior of SFA when applied to high-dimensional visual input data that mimics the visual input of a rat moving in typical experimental environments. The results show that SFA is able to extract abstract information such as the rat's position and head-direction from such complicated input data.

4.1 Nonlinear Blind Source Separation

The problem of blind source separation is simple to state: How can we reconstruct a set of statistically independent sources $\mathbf{s}$ from a set of signals $\mathbf{x} = \mathbf{F}(\mathbf{s})$ that are (possibly nonlinear and convolutive) mixtures of the sources?

This problem has been extensively studied for the case where the signals $\mathbf{x}$ are a linear, invertible and instantaneous mixture $\mathbf{x} = \mathbf{A}\mathbf{s}$ of the sources. The sources can then be reconstructed my means of a linear transformation. Let $\mathbf{y} = \mathbf{W}\mathbf{x}$ with some matrix $\mathbf{W}$ be the estimate of the sources. Then the condition of statistical independence of the estimated sources $\mathbf{y}$ is in general sufficient to constrain the matrix $\mathbf{W}$ up to trivial scaling and permutation transformations. Thus, if we manage to extract statistically independent signals $\mathbf{y}$ from the input signals $\mathbf{x}$, we have essentially solved the problem. For this reason, the problem of linear blind source separation is in essence that of independent component analysis (Hyvärinen et al., 2001).

The problem becomes much harder for nonlinear mixtures. In this case, the inversion of the mixture obviously requires a nonlinear function. Unfortunately, this function is not sufficiently constrained by the requirement that its output signals are statistically independent. Any nonlinear transformation of, say, the first source, is still independent of the other sources, so that there is an infinite number of nonlinear functions that generate statistically independent output signals but fail to recover the sources. Moreover, it has been shown that point nonlinearities of the sources are not the only cause of ambiguities, but that it is even possible to construct nonlinear mixtures of the sources that generate statistically independent output signals. Thus, additional constraints are necessary that distinguish the sources from nonlinearly distorted versions. A discussion of possible

constraints and the resulting approaches to nonlinear BSS can be found in (Jutten & Karhunen, 2003).

Recently, Blaschke et al. proposed that the additional objective of slowness could be used to resolve the ambiguity introduced by the nonlinearity (Blaschke et al., 2007). The argument they put forth is that nonlinearly transformed versions of a signal tend to vary more quickly than the original signal. An intuitive example is the frequency doubling property of a quadratic nonlinearity. Moreover, it was shown by Blaschke et al. (2006) that there is a close relation between SFA and techniques for linear BSS that rely on time-delayed second order statistics of the input signals (TDSEP; Molgedey & Schuster, 1994; Belouchrani et al., 1997; Ziehe & Müller, 1998). A kernel-based nonlinear version of this technique has been shown to successfully recover the sources from rather complicated mixtures (k-TDSEP; Harmeling et al., 2003). The close connection between TDSEP and SFA in combination with the nonlinearity of the SFA algorithm makes SFA an interesting candidate technique for nonlinear BSS problems. However, because SFA extracts not only the sources, but also higher-order harmonics and products thereof, SFA alone is not sufficient for nonlinear BSS. An approach that complements SFA with second order ICA has recently been presented by Blaschke et al. (ISFA; 2007).

Here, we use the theory of chapter 3 to propose a different SFA-based approach to BSS. In section 3.3 we have shown that if we use a nonlinear mixture of statistically independent sources as the input signal for SFA with an unrestricted function space, the sources are represented in the output signals in a very specific fashion. Some of the output signals are predicted to be monotonous point-transformations of the single sources alone, i.e., their lowest non-constant harmonics. For nonlinear BSS, these signals can be considered good representatives of the sources. The other output signals should be higher-order harmonics of single sources or products thereof. Thus, to extend SFA for nonlinear blind sources separation, all we need to do is determine which of its output signals are the lowest harmonics of the sources and discard the other signals. If the sources are Gaussian, the theory predicts that such an approach should even reconstruct the sources exactly.

The simulations presented in this section were done in collaboration with Tiziano Zito.

4.1.1 XSFA: A New Algorithm for BSS

According to the theory in section 3.3, the first output signal found by SFA is the first non-constant harmonic g_{11} of the slowest source. The nature of the second output signal is less clear, however. It can be a higher harmonic of the first source or the first non-constant harmonic of the second source or even a mixture thereof. The idea behind the algorithm we propose here is that once we know the first source, we also know all its possible nonlinear transformations. We can thus remove all aspects of the first source from the SFA output signals by projecting the latter to the space that is uncorrelated to all nonlinear versions of the first source. The remaining signals must depend on the second or even faster sources. The slowest possible signal in this space is then generated by the first non-constant harmonic of the second source, which we can therefore extract by means of linear SFA. Once we know the first two sources, we can proceed by calculating all the harmonics of the second source and all products of the harmonics of the first and the second source and remove those signals from the data. The slowest signal that remains is the first harmonic of the third source. Iterating this scheme should in principle

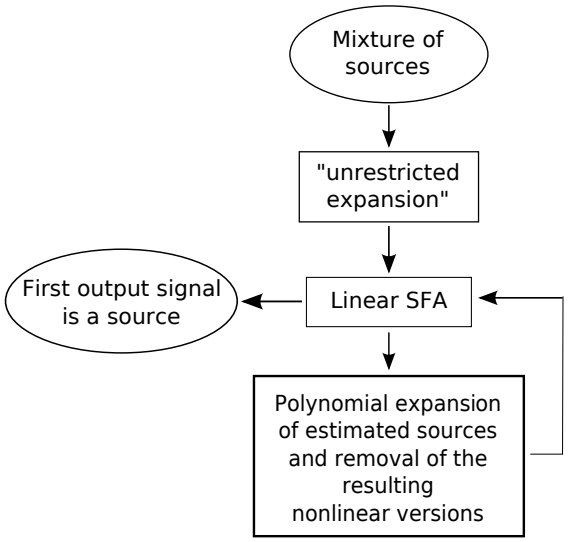


Figure 4.1: **Sketch of the XSFA algorithm.** For simplicity, *sources* and *harmonics* are used synonymously.

yield all the sources.

The structure of the algorithm is illustrated in Figure 4.1. Note that it is a mere extension of SFA in that it does not include new objectives or constraints. We therefore term it XSFA for *eXtended SFA*.

In practice, finite sampling time and restrictions of the function space can complicate the selection process for faster sources. These complications occur when two of the output signals predicted by the theory have approximately the same Δ-value. In this case random correlations corrupt the solution and we get random mixtures of the theoretically predicted solutions. This problem occurs mainly, when the sources have either very similar or very different Δ-values. If they are similar, the algorithm may yield a random (linear) mixture of the sources. This problem can be solved by standard techniques for linear BSS. Because the temporal statistics of the sources is similar, ICA techniques that rely on higher order statistics (Bell & Sejnowski, 1995; Hyvärinen, 1999; Blaschke & Wiskott, 2004) may be favorable over second-oder techniques that rely on temporal correlations (Molgedey & Schuster, 1994; Belouchrani et al., 1997; Ziehe & Müller, 1998). In our simulations, however, the choice of the ICA technique had practically no influence on the performance. The results presented here were obtained using second-order ICA. Note that the situation where the output signals could be a linear mixture of the sources can be detected blindly, because it is sufficient to check if they have similar Δ-values or not.

If the Δ-values of the sources are very different, the algorithm will almost certainly find the first source. However, because the first source is so much slower than the second, the Δ-values of the products $g_{(j,1)} = g_{1j} g_{21}$ of the second source and the harmonics of the first are similar to the Δ-value of the second source g_{21} alone, so that the algorithm may not find the second source, but rather a linear mixture of g_{21} and product solutions $g_{(j,1)}$. In this case, it is not obvious, how the problem can be tackled, because the signals g_{21} and $g_{(j,1)}$ are not statistically independent, so that the usual techniques for linear BSS cannot be expected to disentangle the mixture. In practice, however, second order ICA seems

to solve the problem with more than chance level. At this point, we have no conclusive argument why this is the case. The situation of possible mixtures of product solutions can also be detected blindly, because after projecting out nonlinear versions of the first source and performing linear SFA, the slowest solution is only likely to be a mixture, if the Δ-value of the second slowest output signal is similar.

The simulations presented in the following section apply the following scheme to a nonlinear mixture $\mathbf{x}(t),\ t \in \{1, ..., T\}$ of two sources $\mathbf{s}(t)$:

1. Apply SFA to a polynomial expansion of degree N^{SFA} of the mixture $\mathbf{x}$ and store the J slowest output signals $y_j^{(1)}$.

2. Choose the slowest output signal $y_1^{(1)}$ as a representative $\tilde{s}_1$ of the first source.

3. Expand the representative $\tilde{s}_1$ of the first source in monomials of degree N^{nl} and whiten the resulting signals. We refer to the resulting nonlinear versions of the first source as $n_k,\ k \in \{1, ..., N^{\mathrm{nl}}\}$.

4. Remove the nonlinear versions of the first source from the SFA output signals $y_j^{(1)}$

$$y_j^{(2)}(t) = y_j^{(1)}(t) - \sum_{k=1}^{N^{\mathrm{nl}}} \mathrm{cov}(y_j^{(1)}, n_k)\, n_k(t) \tag{4.1}$$

   and remove directions in which the variance is below a threshold.

5. Apply linear SFA to $y_j^{(2)}$ and store the output signals $y_j^{(3)}$.

6. If the Δ-values of the first output signals $y_1^{(3)}$ and $y_2^{(3)}$ are similar, apply second order ICA to the first components of $y_j^{(3)}$ to disentangle possible linear mixtures of the second source with products of the second source and harmonics of the first. Choose the output signal with the smallest Δ-value as a representative $\tilde{s}_2$ of the second source.

7. If the Δ-values of $\tilde{s}_{11}$ and $\tilde{s}_{21}$ are similar, apply second order ICA to invert possible linear mixtures of the sources.

4.1.2 Simulations

Sources

We evaluated the performance of the algorithm on two different test sets of audio signals. Data set A consists of excerpts from 14 string quartets by Bela Bartok. Note that these sources are from the same CD, the same composer and contain the same instruments. They can thus be expected to have similar statistics. Differences in the Δ-values should mainly be due to short-term nonstationarities. This data set provides evidence that the algorithm is able to distinguish between signals that have similar global statistics based on short-term fluctuations in their statistics.

Data set B consists of 20 excerpts from popular music pieces from various genres, ranging from classical music via rock to electronic music. The statistics of this set is more variable in their Δ-values, in particular they remain different even for long sampling times.

All sources were sampled at 44,100 Hz and 16 bit, i.e., with CD-quality.

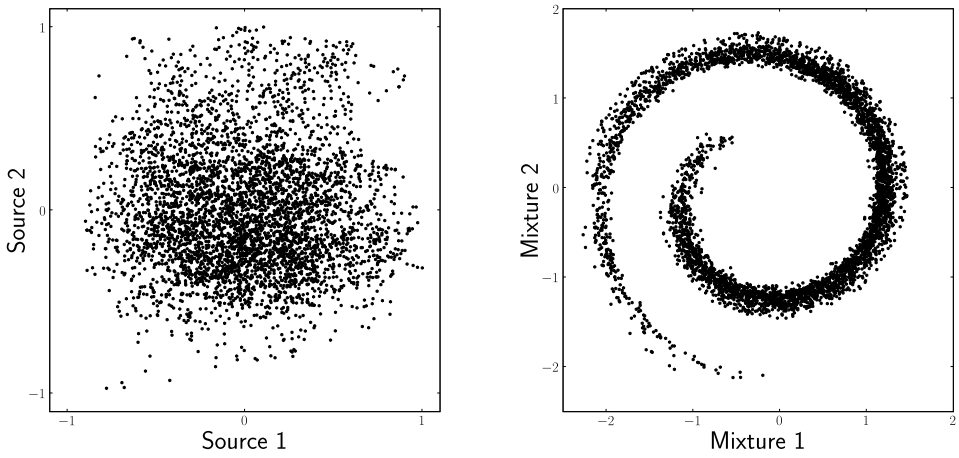


Figure 4.2: **The spiral-shaped structure of the nonlinear mixture.** Panel A shows a scatter plot of two sources from data set A. Panel B show a scatter plot of the nonlinear mixture we used to test the algorithm.

Nonlinear Mixture

We subjected all possible pairs of sources within a data set to a nonlinear mixture that was previously used by Harmeling et al. (2003) and Blaschke et al. (2007):

$$\begin{aligned} x_1(t) &= (s_2(t) + 3s_1(t) + 6)\cos(1.5\pi s_1(t))\,, \\ x_2(t) &= (s_2(t) + 3s_1(t) + 6)\sin(1.5\pi s_1(t))\,. \end{aligned} \tag{4.2}$$

Figure 4.2 illustrates the spiral-shaped structure of this rather extreme nonlinearity. Because this mixture is only invertible if the sources are bounded between -1 and 1, we rescaled the sources to fulfill this condition. The mixture (4.2) is not symmetric in s_1 and s_2. Thus, for every pair of sources, there are two possible mixtures and we have tested both for each source pair.

We have also tested all other nonlinearities used by Harmeling et al. (2003) as well as post-nonlinear mixtures, i.e., linear mixture followed by a point nonlinearity. The performance was similar for all tested mixtures without any tuning of parameters (data not shown). Moreover, the performance remained practically unchanged when we used linear mixtures or no mixture at all. This is in line with the argument that the mixture should be immaterial to SFA if the function space $\mathcal{F}$ is sufficiently rich (see section 3.2).

Simulation Parameters

Degree of the expansion in the first SFA step: We used a polynomial expansion of degree $N^{\mathrm{SFA}} = 7$, because it has previously been shown that this function space is sufficient to invert the mixture (4.2) (Blaschke et al., 2007). For 2-dimensional input signals, this expansion generates a 35-dimensional function space. We kept all $J = 35$ output signals of SFA. It is worth noting that the success rate of the algorithm remains practically unchanged when polynomials of higher order are used. From the theoretical perspective, this is not surprising, because once the function space is sufficiently rich to extract the first non-constant harmonics of the sources, the system performs just as good as it could

with an unrestricted function space.

Degree of the expansion for source removal: We expanded the estimate for the first source in polynomials of degree $N^{\mathrm{nl}} = 20$, i.e., we projected out 20 nonlinear versions of the first source.

Variance threshold: After the removal of the nonlinear versions of the first source, there will be at least one direction with vanishing variance. To avoid numerical problems caused by singularities in the covariance matrices, directions with variance below 10^{-7} were removed. For almost all source pairs, only the trivial direction of the first estimated source was removed.

Parameters for TDSEP steps: We used time delays for the covariance matrices in TDSEP that were equally spaced at 100 samples. The maximal delay was 44100 samples, which corresponds to 441 different time delays within 1s. If the training data were shorter than the maximal delay, the total number of delays was limited by the duration of the training data. For the TDSEP step that should separate the second source and product solutions, we used only those signals whose Δ-values differed by a factor of less than 1.4 from the slowest signal. The final TDSEP step to separate linear mixtures of similarly slow sources was only done, if the extracted signals differed by a factor of less than 1.7 in their Δ-value.

The simulations were done in PYTHON using the modular data processing toolbox (MDP) developed by Berkes & Zito (2007).

Performance Measure

For stationary Gaussian sources, the theory predicts that the algorithm should reconstruct the sources exactly. In most applications, however, the sources will be neither Gaussian nor stationary (at least not on the time scales we used for training). In this case the algorithm cannot be expected to find the sources themselves, but rather their lowest non-constant harmonics.

Thus, the correlation between the output signals of the algorithm and the sources is not necessarily the appropriate measure for the validity of the theory. Therefore, we calculated the lowest non-constant harmonics $g_{\alpha 1}$ of the sources by applying SFA with a polynomial expansion of degree 11 to the individual sources separately and then calculated the correlations between the output signals of the algorithm and both the output signals of the harmonics $y_{\alpha 1}(t) = g_{\alpha 1}(s_\alpha(t))$ and the sources themselves. We considered a source/harmonic to be reconstructed, when the associated correlation was above 0.9.

Simulation Results

Figure 4.3 shows the performance of the algorithm depending on the duration of the training data. For data set A, the algorithm reconstructs the first harmonic of the two sources for roughly 90% of the source pairs for training times longer than 0.2s, corresponding to 88,200 samples. The reconstruction of the sources themselves is equally successful. This may serve as an indication that the sources were close to Gaussian, so that the harmonics and the sources were very similar.

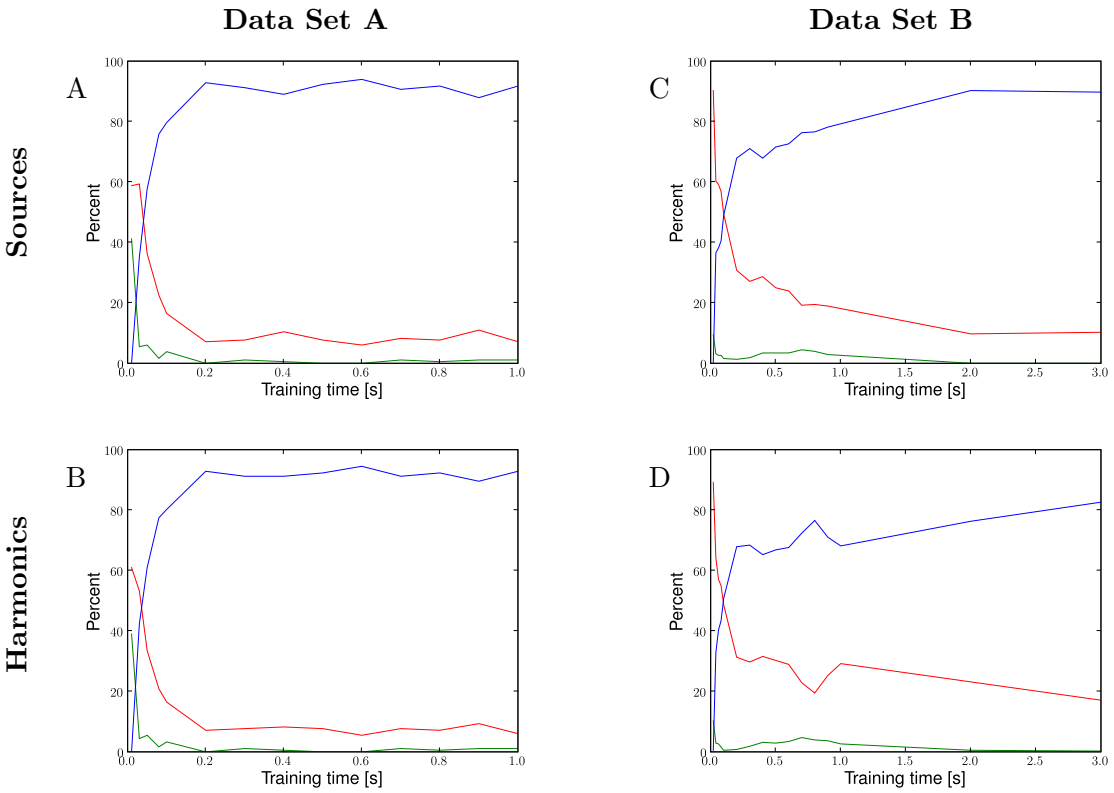


Figure 4.3: **Performance of the algorithm as a function of the duration of the training data.** The curves show the percentage of source pairs, for which the algorithm reconstructed 0 (green), 1 (red) and 2 (blue) of the sources/harmonics. Panels A and B show results for data set A, panels C and D for data set B. Panels A and C show the ability of the algorithm to reconstruct the sources themselves, while B and D show the performance when trying to reconstruct the harmonics of the sources. Statistics cover all possible source pairs that can be simulated (data set A: 14 sources → 182 source pairs, data set B: 20 sources → 380 source pairs).

Similar performance could be achieved for data set B, but longer training times of at least 2s were necessary. Surprisingly, on average, the sources estimated by the algorithm seem to match the original sources better than the harmonics of the sources. A possible reason might lie in the complexity of the function space. If the relation between the harmonics and the sources is highly nonlinear, the function space may be sufficiently complex to find a good approximation of the sources, but not of the harmonics.

To compare the performance of the algorithm with the previously proposed independent slow feature analysis (ISFA) algorithm (Blaschke et al., 2007), we also tested the performance of ISFA on data set A. ISFA is based on a trade-off of a slowness objective Ψ_{SFA} for SFA and an independence objective Ψ_{ICA} that is based on time-delayed second-order correlations:

$$\Psi_{\mathrm{ISFA}} = b_{\mathrm{ICA}}\Psi_{\mathrm{ICA}} - b_{SFA}\Psi_{SFA}\,. \tag{4.3}$$

We fixed the SFA coefficient $b_{SFA} = 1$ and tested four different values of the ICA coefficient: $b_{\mathrm{ICA}} \in \{0, 1, 10, 100\}$. The training data were also sampled at 44,100 Hz and the training duration was 1s, i.e., 44,100 samples. Just as in the original publication, 50 evenly spaced time delays with a maximum of 44,100 samples were used for ISFA.

Number of Reconstructed Sources	ISFA $b_{\mathrm{ICA}} = 0$	ISFA $b_{\mathrm{ICA}} = 1$	ISFA $b_{\mathrm{ICA}} = 10$	ISFA $b_{\mathrm{ICA}} = 100$	XSFA
0	2.8%	40.7%	79.1%	97.8%	1.1%
1	53.8%	52.7%	19.8%	2.2%	7.1%
2	43.4%	6.6%	1.1%	0.0%	91.8%

Table 4.1: **Comparison of the new algorithm and ISFA:** The performance for ISFA and XSFA, tested on the 182 source pairs of data set A. The training time was 1s, i.e., 44,100 samples. ISFA is tested for 4 different values of the trade-off parameter b_{ICA}. Results are given for the reconstruction of the sources themselves, not of the harmonics.

The results are given in table 4.1. Clearly, the average performance drops with increasing values of the ICA coefficient and never reaches the level of the new algorithm. Note that the performance of ISFA is significantly lower than reported by Blaschke et al. (2007)[1]. This discrepancy is due to the simulation procedure: Blaschke et al. used the ideal trade-off of SFA and ICA for each source pair individually. Thus, their data were not obtained in a strictly unsupervised fashion. Moreover, they used a much larger amount of over 2 million training samples.

The performance of XSFA is significantly better that the performance of ISFA as reported in Blaschke et al. (2007) and it is likely that it can be further improved, e.g., by taking more training data or different function spaces.

4.1.3 Practical Limitations of the Theory - Reasons for Failures

There are several reasons why the algorithm can fail, mainly because some of the assumptions underlying the theory are not necessarily fulfilled in simulations. In the following, we discuss some of the reasons for failures.

Limited Sampling Time

The theory predicts that some of the output signals will reproduce the harmonics of the sources exactly. However, problems may arise if output signals have (approximately) the same Δ-value. For example, let us assume that the sources have the same temporal statistics, so that the Δ-value of their slowest harmonics $g_{\alpha 1}$ is equal. Then, there is no reason for SFA to prefer one signal over the other. Rather, any linear mixture of the harmonics $g_{\alpha 1}$ is also an eigenfunction of the operator $\mathcal{D}$.

Of course, in practice, two signals are very unlikely to have exactly the same Δ-value. However, the difference may be so small that it cannot be resolved because of limited sampling length. To get a feeling how well two sources can be distinguished, let us assume that there were only two sources that are drawn independently from probability distributions with Δ-values Δ and $\Delta + \delta$. Then linear SFA should ideally reproduce the sources exactly. However, if there is only a finite amount of data, say a total number of T samples, the Δ-values of the signals can only be estimated with finite precision. Qualitatively, we can distinguish the sources when the standard deviation of the estimated Δ-value is smaller than the difference δ in the "exact" Δ-values. It is clear that this standard deviation will depend on the number of data points roughly as $1/\sqrt{T}$. Thus

[1] They reported that two sources were reconstructed in 70% of the source pairs for a nonlinear expansion of degree 5.

the smallest difference $\delta_{\min}$ in the Δ-values that can be resolved will have the functional dependence

$$\delta_{\min} \sim \Delta^{\alpha} \frac{1}{\sqrt{T}} \,. \tag{4.4}$$

The reason why the smallest distinguishable difference δ must depend on the Δ-value is that neighboring data points are not statistically independent because the signals have a temporal structure. For slow signals, i.e., signals with a small Δ-values, the estimate of the Δ-value will be less precise than for quickly varying signals, because the finite autocorrelation time of the signals impairs the quality of the sampling.

The units for both the Δ-value and $\delta_{\min}$ contain the inverse square of the time unit, while T is measured in time units (in the present context the sampling index acts as a time variable). Thus, for reasons of dimensionality, equation (4.4) requires that the exponent α takes the value $\alpha = 3/4$, yielding the criterion

$$\frac{\delta_{\min}}{\Delta} \sim \frac{1}{\sqrt{T\sqrt{\Delta}}} \,. \tag{4.5}$$

For an interpretation of this equation note that the Δ-value can be interpreted as a (quadratic) measure for the width of the power spectrum of a signal:

$$\Delta(y) = \frac{1}{T} \int \dot{y}^2 \mathrm{d}t = \frac{1}{T} \int \omega^2 |y(\omega)|^2 \mathrm{d}\omega \,, \tag{4.6}$$

where $y(\omega)$ denotes the Fourier transform of $y(t)$. However, the inverse width of the power spectrum is an operative measure for the autocorrelation time τ of the signal, leaving us with $\tau \sim 1/\sqrt{\Delta}$. With this in mind, the criterion (4.5) takes a form that is much easier to interpret:

$$\frac{\delta_{\min}}{\Delta} \sim \sqrt{\frac{\tau}{T}} = \frac{1}{\sqrt{N_\tau}} \,. \tag{4.7}$$

τ characterizes the time scale on which the signal varies, so intuitively, we can cut the signal into $N_\tau = T/\tau$ "chunks" of duration τ, which are approximately independent. The estimate (4.7) then states that the smallest relative difference in the Δ-value that can be resolved is inversely proportional to the square root of the number N_τ of independent data "chunks".

If the difference in the Δ-value of the predicted solutions is smaller than $\delta_{\min}$, SFA is likely not to find the predicted solutions but rather an arbitrary mixture thereof, because the removal of random correlations and not slowness will be the essential determinant for the solution of the optimization problem. The relation (4.7) may serve as a rough guideline for how much training time is needed to distinguish two signals. Note however, that the validity of (4.7) is questionable for nonstationary sources, because the statistical arguments used above are not valid. This may be one of the reasons, why data set B requires more training data, because some of these sources contain percussive instruments that can generate very variable statistics, at least on the sub-second time scale (periods of low amplitude, interspersed with short periods of high amplitude and frequency associated, e.g., with cymbals).

Density of Eigenvalues

The problem of getting random mixtures instead of the optimal solutions is most obvious in the case where the sources, or more precisely, the slowest non-constant harmonics of the

sources, have similar Δ-values. However, even when the sources are sufficiently different, this problem will eventually arise for the higher-order solutions. To quantify the expected differences in Δ-value between the solutions, we define a *density*[2] $\rho(\Delta)$ of the Δ-values as the number of eigenvalues expected in an interval $[\Delta, \Delta + \delta]$, divided by the interval length δ. A convenient way to determine this density is to calculate the number $R(\Delta)$ of solutions with eigenvalues smaller than Δ and then take the derivative with respect to Δ.

In the Gaussian approximation, the Δ-values of the harmonics are equidistantly spaced, that is $\lambda_{\alpha i_\alpha} = i_\alpha \lambda_{\alpha 1}$. As the Δ-value $\Delta_{\mathbf{i}}$ of the full product solution $g_{\mathbf{i}}$ is the sum of the Δ-values of the harmonics, the condition $\Delta_{\mathbf{i}} < \Delta$ restricts the index $\mathbf{i}$ to lie below a hyperplane with the normal vector $\mathbf{n} = (\lambda_{11}, ..., \lambda_{S1}) \in \mathbb{R}^S$, because

$$\Delta_{\mathbf{i}} \overset{(3.49)}{=} \sum_\alpha i_\alpha \lambda_{\alpha 1} = \mathbf{i} \cdot \mathbf{n} < \Delta \,. \tag{4.8}$$

Since the indices are homogeneously distributed with spacing 1 in index space, the expected number of solutions with $\Delta < \Delta_0$ is simply the volume of the subregion in index space for which equation (4.8) holds:

$$R(\Delta) = \frac{1}{S!} \prod_{\alpha=1}^{S} \frac{\Delta}{\lambda_{\alpha 1}} \,. \tag{4.9}$$

The density of the eigenvalues is then given by

$$\rho(\Delta) = \frac{\partial R(\Delta)}{\partial \Delta} = \frac{1}{(S-1)!} \left[\prod_\alpha \frac{1}{\lambda_{\alpha 1}} \right] \Delta^{S-1} \,. \tag{4.10}$$

As the density of the eigenvalues can be interpreted as the inverse of the expected distance between the Δ-values, the distance and thus the separability of the solutions with a given amount of data will decline as $1/\Delta^{S-1}$. In simulations, we can thus expect to find the theoretically predicted solutions only for the slowest functions, higher order solutions will tend to be linear mixtures of the theoretically predicted functions. This is particularly relevant if there are many sources, i.e., when S is large.

If the sources are not Gaussian, the dependence of the density on the Δ-value can differ (e.g., for uniformly distributed sources[3], $\rho(\Delta) \sim \Delta^{S/2-1}$). The problem of decreasing separability, however, will generally remain, at least when the number of sources is not extremely small.

Sampling Rate

The theory is derived under the assumption that all signals are continuous in time. Real data will always be discretized. Therefore, the theory will only be valid if the data are sampled with a sampling rate that is sufficient to generate quasi-continuous data. As the sampling rate decreases, so will the correlations between subsequent data points, rendering techniques like SFA that are based on short-term temporal correlations useless.

[2] In statistical and condensed matter physics, the quantity ρ is usually referred to as the *density of states* (Kittel et al., 1986).

[3] For uniformly distributed sources, the Δ-value depends quadratically on the spatial frequency/index of the solution. Using the analogy to physics (see section 3.4), the density ρ of eigenvalues is thus given by the density of states for a quadratic dispersion relation (see, e.g., Kittel et al., 1986).

For extremely low sampling rates, the signals effectively become white noise. In this case any nonlinear transformation that generates a signal with unit variance would have the same Δ-value (i.e., $\Delta = 2$) and SFA would generate a set of random nonlinear mixtures of the signals.

But what happens in the intermediate case, when the data still contain significant temporal correlations, but cannot be considered quasi-continuous? To address this question, we have to consider that the Δ-values of the discretely sampled sources are bounded by $\Delta = 4$ from above and by the Δ-value of the first harmonic from below. Thus, for ill-sampled sources, the Δ-values of the harmonics become more similar than for well-sampled sources, because they are "quenched" into a smaller interval. Since functions with similar Δ-values tend to mix more easily in the presence of random correlations, we expect that the slowest function found by SFA is not necessarily the first harmonic, but that it may contain components of higher harmonics. Then, the monotonous relation between the sources and their estimate is no longer ensured, so that XSFA may extract unwanted nonlinear transformations of the sources instead of the sources themselves.

Note that low sampling rates lead to large Δ-values, which can thus serve as a first indication if the sampling rate is sufficient.

Function Space

An assumption of the theory is that the function space $\mathcal{F}$ accessible to SFA is unlimited, but any application has to restrict the function space to finite dimension. If the function space is ill-chosen in that it cannot invert the mixture that generated the input data from the sources, it is clear that the theory can no longer be valid.

Because the nature of the nonlinear mixture is not known a priori, it is difficult to choose an appropriate function space. We used polynomials with relatively high degree. A problem with this choice is that polynomials of high degree generate extremely sparse data distributions, i.e., the majority of data points lies around zero with few very large exceptions. Depending on the input data at hand, it may be more robust to use other basis functions such as radial basis functions. To increase the dimensionality of the functions space one can also use kernel methods (Bray & Martinez, 2002).

The suitability of the function space is one of the key determinants for the quality of the estimation of the first source. If this estimate is not accurate but has significant contributions from other sources, the nonlinear versions of the estimate that are projected out will not be accurate, either. The projection step may thus remove aspects of the second source and thereby impair its estimate. We expect that for many sources, these errors will accumulate so that estimates for faster sources will not be trustworthy. This problem might be further complicated by the increasing eigenvalue density discussed above.

4.1.4 Discussion

Here, we have used the theory of chapter 3 to propose and test a new algorithm for nonlinear blind source separation. The basis of the algorithm is the analytical prediction that SFA represents statistically independent sources in terms of products of their harmonics and that some of these harmonics should be monotonic functions of the sources themselves. XSFA uses this to iteratively reconstruct the sources, in theory from arbitrary invertible mixtures. Simulations for a rather complicated nonlinear mixture of

two audio sources have shown that the algorithm extracts both sources for 90% percent of the source pairs. The performance is substantially higher than the performance of independent slow feature analysis (ISFA; Blaschke et al., 2007), another algorithm for nonlinear BSS that relies on temporal correlations.

An important advantage of the new algorithm over ISFA is that it is robust to changes of the implementation details. Neither a higher degree of the expansion before the first SFA step nor the removal of more nonlinear versions of the first source change the reconstruction performance significantly. It should be noted, however, that polynomial expansions - as used here - become problematic if the degree of the expansion is too high. The resulting expanded data contain directions with very sparse distributions, which can lead (a) to singularities in the covariance matrix (e.g., for Gaussian signals x with limited sampling, monomials of high and even order (e.g., x^{20} and x^{22}) are almost perfectly correlated) and (b) to sampling problems for the estimation of the required covariances because the data are dominated by few data points with high values. Note, that this problem is not specific to the algorithm itself, but rather to the expansion type used. Other expansions such as radial basis functions may be more robust. The relative insensitivity of XSFA to parameters is a major advantage over ISFA, because an algorithm that requires fine-tuning of parameters depending on the sources at hand is, strictly speaking, not unsupervised.

Many algorithms for nonlinear BSS are designed for specific types of mixtures, e.g., for post-nonlinear mixtures (for an overview of methods for post-nonlinear mixtures see Jutten & Karhunen (2003)). In contrast, our algorithm should work for arbitrary instantaneous mixtures. As previously mentioned, we have performed simulations for a set of instantaneous nonlinear mixtures, which yielded similar performance. The only requirements are that the sources are distinguishable based on their temporal statistics and that the function space accessible to SFA is sufficiently complex to invert the mixture. Note that the algorithm is restricted to instantaneous mixtures. It cannot invert convolutive mixtures because SFA processes its input instantaneously and is thus not suitable for deconvolution tasks.

We have presented simulations for two sources only. In theory, the algorithm should be able to separate mixtures of more sources as well. In practice, however, the number of reconstructable sources may be limited because of accumulating errors as discussed in section 4.1.3. Further simulations will be needed to assess the performance of the algorithm for more sources.

In summary, we have presented a new algorithm for the complicated problem of nonlinear blind source separation that is (a) completely unsupervised, (b) independent of the mixture, (c) robust to parameters, and (d) reliable (for a nonlinear BSS technique), as shown by the reconstruction performance of about 90% for the examined case of two audio sources. Moreover, the algorithm is underpinned by a rigorous mathematical framework, which is not the case for most other BSS algorithms.

4.2 Place and Head-Direction Codes

In the last section, we have provided computational evidence that the theory developed in chapter 3 can describe the behavior of SFA when applied to an arbitrary nonlinear mixture of statistically independent sources. The independence of the performance on the mixture indicates that the mixture can indeed be understood as a coordinate transformation that is immaterial to the algorithm, as discussed in section 3.2. There, we also argued that there should be no difference if the input signals are high-dimensional embeddings of a low-dimensional manifold or if they are a low-dimensional parameterization of the manifold. In this section, we test this hypothesis by applying SFA to a self-localization task based on high-dimensional visual input.

All simulation results presented in this chapter have been performed by Mathias Franzius. The results presented in this section were published in (Franzius, Sprekeler & Wiskott, 2007).

4.2.1 The Problem of Self-Localization

In order to successfully navigate in our environment, we have to know where we are and which direction we are heading. To retrieve this information, our brain has access to two types of information. Firstly, there are internal cues such as acceleration or velocity signals transmitted from the vestibular system or proprioceptive sensors. By temporal integration of this information, we can infer where we are, given we knew where we were at an earlier moment in time. This process is usually referred to as path integration or dead reckoning. The advantage of path integration is that it works even in the absence of sensory stimuli, e.g., in darkness. The disadvantage is that it tends to accumulate errors, so that the inferred position becomes less and less precise with time. To correct for these errors, a second category of information is needed: sensory stimuli.

In familiar environments, sensory information is usually sufficient for us to determine where we are. We have to know the environment, however, so this form of self-localization is an acquired ability, a result of learning. Here, we argue that slowness learning, and SFA in particular, may be an appropriate mechanism for learning to self-localize purely based on visual information.

Let us consider a rat exploring a room that remains perfectly static over time. The set of images it can perceive in such a setup forms a small subset of all possible images: Its visual input is uniquely determined by its position and the direction of its gaze. Hence, the set of possible images forms a low-dimensional manifold in the high-dimensional space of all images.

If we use the visual input that the rat perceives during the exploration of the room as input signals for an SFA system with an unrestricted function space, we are facing a system that fulfills all conditions for the theory developed in chapter 3. In particular, the output signals should not depend on whether we use the images as input data or abstract coordinates that specify the rat's spatial configuration. In the following, we will use the theory to make analytical predictions for the dependence of the output signals of SFA on the rat's position and head-direction. We will concentrate the considerations on two scenarios that are commonly used in the laboratory: experiments in open fields and linear tracks. The predictions are compared with simulation results as published in (Franzius, Sprekeler & Wiskott, 2007). The theory and the simulations are in excellent agreement and show that the output of SFA reflects the spatial coordinates of the rat

in an orderly fashion. This makes SFA an interesting element for models of spatial navigation in animals (Franzius, Sprekeler & Wiskott, 2007) as well as for navigation tasks in robotics.

It is interesting to note that if we make the additional assumption that the coordinates characterizing the rat's spatial position are mutually independent, e.g., that its position has no influence on the direction in which it is looking, we can interpret the self-localization task as a nonlinear blind source separation problem: A set of statistically independent sources (position and head-direction) are nonlinearly mixed into a high-dimensional representation (the images the rat perceives). If the visual structure of the room is sufficiently complex, the rat can infer its position and head-direction from the visual input, so this mixture is in general invertible. As predicted by the theory in section 3.3, the output of the SFA units can then be factorized into functions that depend on one of the coordinates only.

In the article (Franzius, Sprekeler & Wiskott, 2007) we presented a hierarchical network that learns place cells, head-direction cells and spatial view cells from complicated visual input. The network consists of two basic components. First, SFA is applied to visual data to learn invariant, but distributed representations of the spatial coordinates we are interested in. Then, an additional layer of linear sparse coding transforms these distributed representations into localized representations that resemble those found in the brain. Here, we will focus on the SFA component of the network, because it allows the application of the theory developed in chapter 3. We will use the simulation results mainly to illustrate that the theory can describe the simulations and to give a qualitative discussion of the representation after the sparse coding step. For this reason, we will skip most technical details of the simulations and refer the interested reader to (Franzius, Sprekeler & Wiskott, 2007) and to the PhD thesis of Mathias Franzius (Franzius, 2008).

4.2.2 Open Field Experiments

Ever since the discovery of so-called place cells in the rodent hippocampus (O'Keefe & Dostrovsky, 1971), neural correlates of spatial navigation have become a major topic in neuroscientific research. This trend was further strengthened by the discovery of head-direction cells (Taube et al., 1990) and, recently, grid cells (Hafting et al., 2005) in the rodent hippocampal formation.

A typical experimental setup is the so-called *open field* experiment: a rat that can move freely within an enclosure is frequently fed with food pellets, which are dropped at random locations. Usually, the rat's position and possibly also its head direction are monitored with a camera and correlated with physiological data, typically extra-cellular recordings.

Rectangular Open Field

In open field experiments, the spatial configuration of the rat is usually parameterized by its position, indicated by the coordinates x and y, and its head direction ϕ. It is common to neglect additional degrees of freedom such as the vertical angle of the rat's gaze (pitch). For rectangular arenas, the configuration of the rat can thus be characterized by a *configuration vector* $\mathbf{s} = (x, y, \phi) \in [0, L_x] \times [0, L_y] \times [0, 2\pi[$, where L_x and L_y denote the size of the room in x- and y-direction, respectively. We choose the origin of the head direction ϕ such that $\phi = \pi/2$ corresponds to the rat looking to the North and

increasing ϕ results in clockwise rotation, e.g., from North towards East. The dynamics of the rat's motion can be characterized by a (generalized) velocity vector $\mathbf{v} = (v_x, v_y, \omega)$, where v_x and v_y denote the translation velocities and ω is the angular rotation velocity. In the typical pellet-throwing experiment, the rat shows random search behavior, so it is reasonable to assume that the velocities in the three different directions are uncorrelated and that the rat's position and head direction are homogeneously distributed. Moreover, the variance of the velocity should be similar in x- and y-direction. The covariance matrix of the velocities then takes the form

$$\mathbf{K} = \begin{pmatrix} \langle v^2 \rangle & 0 & 0 \\ 0 & \langle v^2 \rangle & 0 \\ 0 & 0 & \langle \omega^2 \rangle \end{pmatrix} \tag{4.11}$$

and the probability density $p(x, y, \phi)$ is a constant. Note that due to anatomical constraints, the head-direction and the movement velocity should be correlated, because the rat tends to move towards the direction it it facing. For simplicity, we will neglect this correlation. It would introduce a dependence of $K_{\mu\nu}$ on head-direction. Although the movement statistics used for the simulations included this correlation, the theoretical predictions agree with the simulation results, as the reader will see below.

In this case the eigenvalue problem (3.23) for the optimal functions for SFA becomes

$$-\left[\langle v^2 \rangle \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \langle \omega^2 \rangle \frac{\partial^2}{\partial \phi^2}\right] g(x, y, \phi) = \Delta g(x, y, \phi)\,. \tag{4.12}$$

The boundary conditions (3.24) yield

$$\frac{\partial}{\partial x} g(x, y, \phi) = 0 \quad \text{for} \quad x \in \{0, L_x\}, \tag{4.13}$$

$$\frac{\partial}{\partial y} g(x, y, \phi) = 0 \quad \text{for} \quad y \in \{0, L_y\}, \tag{4.14}$$

and cyclic boundary conditions in the angular direction.

It is easy to check that the eigenfunctions and the corresponding Δ-values are given by

$$g_{lmn}(x, y, \phi) = \begin{cases} \sqrt{8} \cos(l\pi \frac{x}{L_x}) \cos(m\pi \frac{y}{L_y}) \sin(\frac{n+1}{2}\phi) & \text{for n odd} \\ \sqrt{8} \cos(l\pi \frac{x}{L_x}) \cos(m\pi \frac{y}{L_y}) \cos(\frac{n}{2}\phi) & \text{for n even} \end{cases} \tag{4.15}$$

$$\Delta_{lmn} = \begin{cases} \pi^2 \langle v^2 \rangle \left(\frac{l^2}{L_x^2} + \frac{m^2}{L_y^2}\right) + \langle \omega^2 \rangle \frac{(n+1)^2}{4} & \text{for n odd} \\ \pi^2 \langle v^2 \rangle \left(\frac{l^2}{L_x^2} + \frac{m^2}{L_y^2}\right) + \langle \omega^2 \rangle \frac{n^2}{4} & \text{for n even,} \end{cases} \tag{4.16}$$

with l, m, and n being non-negative natural numbers. Only $l = m = n = 0$ is not allowed, as this case corresponds to the constant solution, which violates the unit variance constraint.

The solutions factorize into three sinusoidal functions, each of which depends on one of the coordinates x, y, and ϕ only. This is not surprising, because for a homogeneously sampled rectangular room, the three coordinates x, y, and ϕ are statistically independent, so that the problem can be understood as a nonlinear blind sources separation problem and the theory for statistically independent sources of section 3.3 applies. Note that the position variables x and y are not statistically independent for arbitrary shapes of the

room. For example, in a circular open field, a fixed value of the x-coordinate introduces constraints for y. Thus, x contains information about y, so the coordinates are not statistically independent. For circular rooms, a change to polar coordinates reestablishes the statistical independence of the coordinates and thus the factorization of the solutions.

To check if the theory is capable of describing the results of numerical simulations, we compare the theoretical prediction (4.15) with the simulation results published in (Franzius, Sprekeler & Wiskott, 2007). The input data for the simulations are high-dimensional, quasi-natural image sequences that show the visual input of a simulated rat in a rectangular virtual-reality environment. The motion of the rat is modeled as a modified Brownian motion with momentum. We used the resulting video sequences to train a hierarchical network of SFA modules, which implements a subset of all polynomials of degree 8 in all input variables. Further technical details of the simulations can be found in (Franzius, Sprekeler & Wiskott, 2007; Franzius, 2008).

To compare the solutions (4.15) with the outcome of the simulations, we need to order them by their Δ-values. For better comparability with the simulations it is convenient to rewrite the Δ-values in the following form:

$$\Delta_{lmn} = \frac{\pi^2 \langle v^2 \rangle}{L_x^2} \times \begin{cases} l^2 + \frac{L_x^2}{L_y^2} m^2 + {v_{\text{rel}}}^2 (n+1)^2 & \text{for n odd} \\ l^2 + \frac{L_x^2}{L_y^2} m^2 + {v_{\text{rel}}}^2 n^2 & \text{for n even,} \end{cases} \tag{4.17}$$

where

$${v_{\text{rel}}}^2 = \frac{\langle (\frac{\omega}{2\pi})^2 \rangle}{\langle (\frac{v}{L_x})^2 \rangle} \tag{4.18}$$

denotes the relative rotational speed, i.e., the ratio of the mean squared rotational and translational velocity, if translational velocity is measured in units of the room size in x-direction per second and rotational velocity is measured in full circles per second.

We can now discuss two limit cases in terms of the relative velocity v_{rel}. Let us first consider the case where the rat moves at small velocities while making a lot of quick turns, i.e., $v_{\text{rel}} \gg 1$. In this case, the smallest Δ-values can be reached by setting $n = 0$, unless $l^2 + \frac{L_x^2}{L_y^2} m^2 > 4 {v_{\text{rel}}}^2$. Since for $n = 0$ the functions g_{lmn} do not depend on the angle ϕ, the slowest functions for this case are invariant with respect to head direction and encode the rat's position. The structure of the solutions (4.15) and the respective simulation results are depicted in panels A and B of Figure 4.4. Panel C shows that by applying sparse coding, the resulting representation of the animals position can be transformed into a localized representation that resembles place fields as found in the hippocampus. For a detailed discussion see below.

In the other extreme, v_{rel} is much smaller than one, i.e., the rat runs relatively fast while making few or slow turns. The smallest Δ-values can then be reached by choosing $l = m = 0$ unless $n^2 > \min(1, \frac{L_x^2}{L_y^2}) / {v_{\text{rel}}}^2$. The corresponding functions are invariant with respect to position while being selective to head direction. A comparison of these theoretically predicted functions with simulation results are shown in panels D and E of Figure 4.4.

The theoretical predictions are in good agreement with the simulation results, both for high and for low relative rotational velocity v_{rel}.

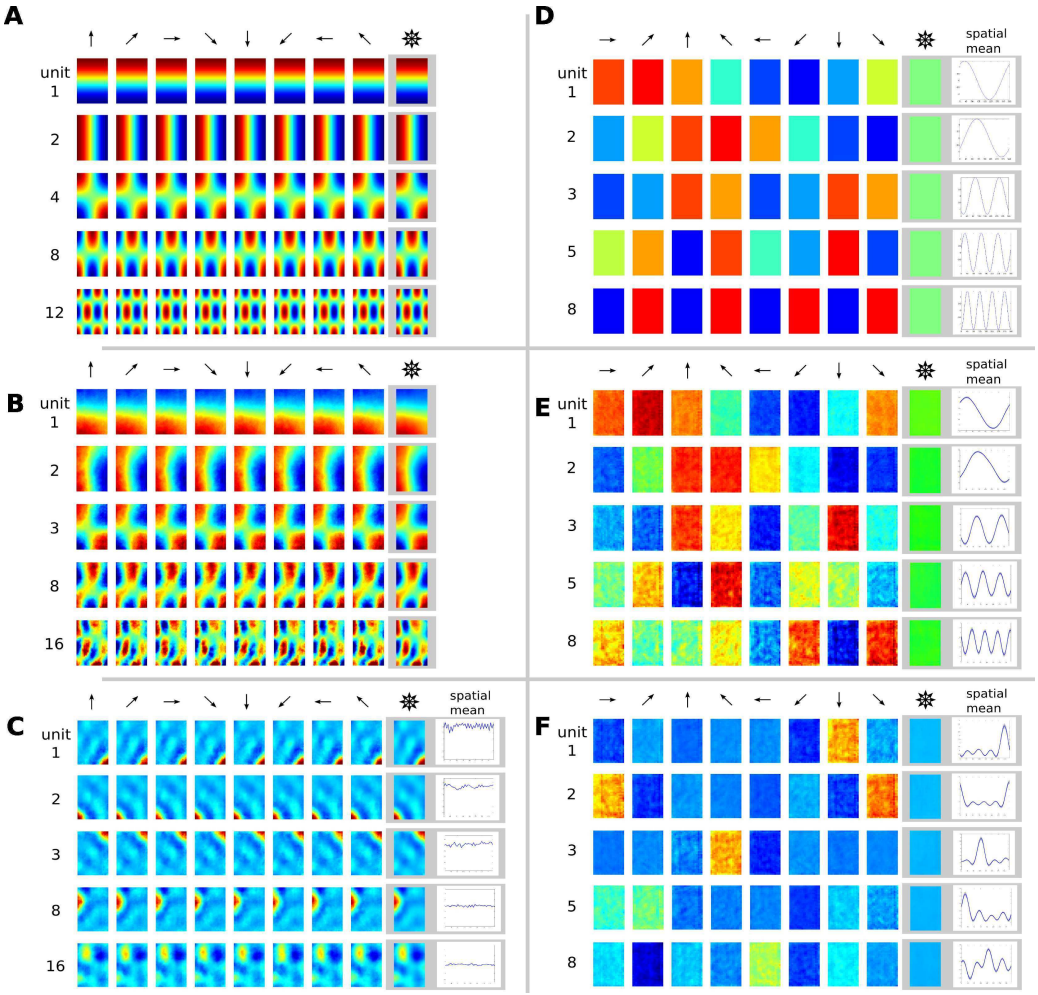


Figure 4.4: **Theoretical Prediction and Simulation Results for the Open Field Experiment.** Each row within each panel shows the response of one SFA unit as a function of position for different head directions (indicated by arrows), as well as the mean value averaged over all head directions (indicated by superimposed arrows). Blue denotes low output values, green intermediate, and red large output values. Panels (C), (D), (E) and (F) also show head direction tuning curves averaged over all positions ± one standard deviation. **(A)** Theoretical prediction for the functions learned by SFA with relatively quick rotation velocity compared with translation velocity ($v_{\text{rel}} = 32$). Solutions are ordered by slowness and have regular grid structures. **(B)** Simulation results for SFA with the same parameters as in (A), ordered by slowness. The results are similar to the theoretical predictions up to mirroring, sign, and mixing of solutions with similar Δ-values. All units are head-direction invariant and code for spatial position. **(C)** Simulation results after sparse coding of the output signals in (B), ordered by sparseness (kurtosis). Sparse coding was implemented by means of cumulant-based independent component analysis (CuBICA, Blaschke & Wiskott, 2004). Output patterns of all units are localized and head-direction invariant, resembling hippocampal place cells. **(D)** Theoretical prediction for the functions learned by SFA for relatively slow rotation velocity and fast translation velocity. All solutions are position-invariant and constitute a Fourier basis in head-direction space. As the phases of the theoretical solutions are not uniquely determined, they were adjusted to match the simulation results in (E). **(E)** Simulation results for SFA for the same settings as in (D) ($v_{\text{rel}} = 0.08$), ordered by slowness. The results are similar to the theoretical predictions. All units are position-invariant and head-direction specific but not localized in head-direction space, i.e., all units except 1 and 2 have multiple peaks. **(F)** Simulation results after sparse coding of the output signals illustrated in (E), again ordered by sparseness (kurtosis). Firing patterns of all units are position invariant and localized in head-direction space, resembling subicular head-direction cells.

4.2.3 Linear Track

Another common paradigm for place cell experiments is the *linear track*, a rectangular enclosure that is very narrow in one direction, so that the (translational) movement of the rat is essentially restricted to one dimension.

In principle, the set of possible spatial configurations for the linear track is the same as for the open field, only with a small side length L_x in one direction. Equation (4.16) shows that for small L_x the solutions that are not constant in the x-direction, i.e., the solutions with $l \neq 0$, have large Δ-values and thus vary quickly. Therefore, slow functions will be independent of x, so we can neglect this dimension and restrict the configuration space to position in y-direction and head direction ϕ.

Another difference between the simulation setup for the open field and the linear track lies in the movement statistics of the rat. In the simulations and in most experimental paradigms, the rat rarely turns on mid-track, but traverses the track from one end to the other. For anatomical reasons, the head direction of the rat is restricted by the direction in which it runs. In the simulations, we modelled this is the following fashion: Given the sign of the velocity in y-direction the head direction is restricted to angles between either 0 and π (positive velocity in y-direction, North) or between π and 2π (negative velocity in y-direction, South). If, in addition, the rat makes a lot of quick head rotations, the optimal functions for SFA can only be slowly varying if they are invariant with respect to head direction within these ranges. For the theory, we can thus consider a reduced configuration space that contains the position y and a binary value $d \in \{\text{North, South}\}$ that determines whether $0 \leq \phi < \pi$ or $\pi \leq \phi < 2\pi$.

Note that this simplified configuration space is also widely used in linear track experiments. One advantage of the linear track is that the remaining dimensions are experimentally much easier to sample smoothly than the full three dimensional parameter space of the open field.

We assume that the rat only switches between North and South at the ends of the track. Because discontinuities in the functions lead to large Δ-values, slow functions $g(y, d)$ should fulfill the continuity condition that $g(0, \text{North}) = g(0, \text{South})$ and $g(L_y, \text{North}) = g(L_y, \text{South})$. This means that the configuration space has the topology of a circle, where one half of the circle represents all positions with the rat facing North and the other half the positions with the rat facing South. It is thus convenient to introduce a different variable $\xi \in [0, 2L_y]$ that labels the configurations in the following way:

$$(x(\xi), d(\xi)) = \begin{cases} (\xi, \text{North}) & \text{for} \quad \xi < L_y \\ (2L_y - \xi, \text{South}) & \text{for} \quad \xi \geq L_y \end{cases} . \tag{4.19}$$

The topology of the configuration space is then captured by cyclic boundary conditions for the functions $g(\xi)$.

For simplicity we assume that there are no preferred positions or head directions, i.e., that both the variance of the velocity $K = \langle \dot{\xi}^2 \rangle$ and the probability distribution $p(\xi)$ is independent of ξ. The equation for the optimal function is then given by

$$- \langle \dot{\xi}^2 \rangle \frac{\partial^2}{\partial \xi^2} g(\xi) = \Delta g(\xi) \, . \tag{4.20}$$

The solutions that satisfy the cyclic boundary condition and their Δ-values are given by

$$g_j(\xi) = \begin{cases} \sqrt{2}\sin(j\pi\frac{\xi}{2L_y}) & \text{for j even} \\ \sqrt{2}\cos((j+1)\pi\frac{\xi}{2L_y}) & \text{for j odd} \end{cases}, \tag{4.21}$$

$$\Delta_j = \begin{cases} \pi^2\frac{\langle\dot{\xi}^2\rangle}{4L_y^2}j^2 & \text{for j even} \\ \pi^2\frac{\langle\dot{\xi}^2\rangle}{4L_y}(j+1)^2 & \text{for j odd} \end{cases}. \tag{4.22}$$

Note that there are always two functions with the same Δ-value. Theoretically, any linear combination of these functions has the same Δ-value and is thus also a possible solution. In the simulation, this degeneracy does not occur, because mid-track turns do occur occasionally, so those functions that are head-direction-dependent on mid-track (i.e., those with even j) will have higher Δ-values than theoretically predicted. This avoids mixed solutions and changes the order of the functions when ordered by slowness. Panel A of Figure 4.5 shows seven of the theoretically predicted functions g_j, reordered such that they match the experimental results shown in panel B. Again, the predictions are in excellent agreement with the simulations (except for the oder).

4.2.4 Place Cells, Grid Cells, and Head-Direction Cells

So far, we have shown that SFA is able to extract high-level information such as an animal's position and head-direction from the visual data it receives and that this information is represented in a very specific way. Bearing in mind that the motivation for SFA is biological in nature, an obvious question is if the representations found by SFA are similar to representations found in the brain.

In the last decades, a range of cell types has been identified that are thought to form neural correlates of self-localization and spatial navigation in rodents. *Place cells* - neural correlates of a rodent's position - were found more than 35 years ago in hippocampal areas CA1 and CA3 (O'Keefe & Dostrovsky, 1971), correlates of head orientation - termed *head-direction cells* - were found 20 years later (Taube et al., 1990), and recently, non-localized representations - termed *grid cells* - were found in entorhinal cortex of rats (Hafting et al., 2005). Primates possibly also have place cells, certainly head-direction cells, and also *spatial-view cells* that do not encode the animal's own position but fire whenever the animal views a certain part of the environment (Rolls, 1999; Horton & Adams, 2005; Rolls, 2006; O'Keefe, 2007).

All of these cells selectively encode certain aspects of position and/or head-direction of the animal while being invariant with respect to others. Head-direction cells are strongly selective for the orientation of the animal's head and largely invariant to its position (Horton & Adams, 2005). They typically have a single peak of activity with a Gaussian or triangular shape and a tuning width of roughly 60° to 150° (Taube & Bassett, 2003). Thus, they share the invariance property with the SFA units for fast translation velocities, but lack the periodicity of the SFA representation of head-direction. In contrast to head-direction cells, most place cells recorded in open fields are invariant to head direction while being selective for the animal's position (Muller et al., 1994). The SFA units for high rotation velocities share these properties, but lack the localized structure of the high activity region. Interestingly, the degree of orientation-invariance of place cells depends on the behavioral task of the animal and possibly on the structure of the environment. In linear tracks, and for repeated linear paths in open environments most place cells are orientation-specific (Markus et al., 1995).

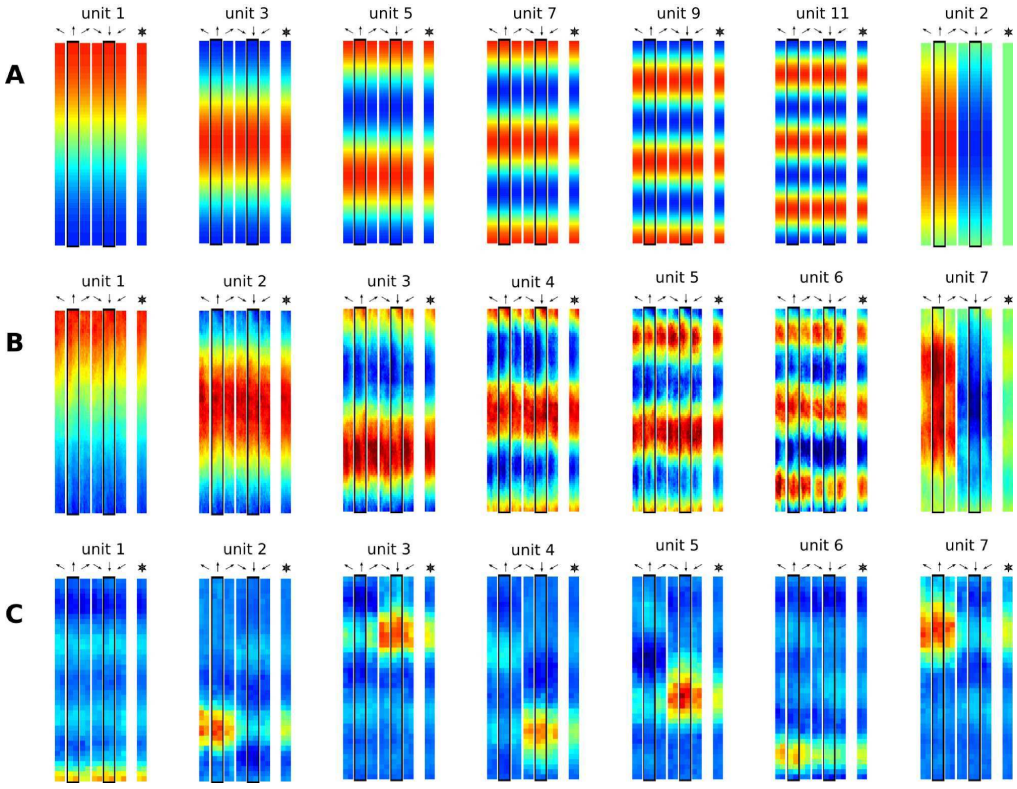


Figure 4.5: **Theoretical Predictions and Simulation Results for the Linear Track.** Head directions are indicated by arrows, orientation averages are indicated by superimposed arrows, and principal directions (North, South) are emphasized with a dark border. Blue color denotes small output, green intermediate, and red large output signals. **(A)** Theoretical predictions, reordered to match the simulation results. **(B)** Spatial output pattern of the first (i.e., the slowest) seven out of ten simulated SFA units. Units 1-6 are mostly head-direction invariant whereas unit 7 responds differently to North and South views. In total, 10 units were simulated. Two out of the three remaining units are also head-direction invariant. **(C)** Spatial activity maps of the first (i.e., most kurtotic) seven out of ten units learned by sparse coding on the output signals of the SFA units in panel B. Sparse coding was implemented by cumulant-based ICA (CuBICA; Blaschke & Wiskott, 2004). All units are localized in space and most of them are only active for either North or South views, closely resembling place fields recorded from rats in linear track experiments.

It is worth noting that the invariances encountered in place and head-direction cells are highly non-trivial. Sensory stimuli, visual stimuli in particular, are in general very different for different positions and – probably even more so – for different head directions. For this reason, self-organized formation of invariant spatial representations based on visual stimuli is a difficult task. This is why many sensory-driven models of place and head-direction cells are based on abstract input data, e.g., on distances and bearings of landmarks (see e.g., Sharp, 1991). The invariant representations learned by SFA are thus an important step towards a model of sensory-driven place and head-direction cells. Indeed, the only feature that distinguishes the SFA representations in open fields from place and head-direction cells is that their responses are not localized.

The lack of localization of the response pattern is a recurring problem when using SFA as a model for neural response properties (see the model for complex cell receptive fields in chapter 5). Interestingly, however, the periodicity of the spatial response pattern

of the SFA units for quick rotations (panels A and B of Figure 4.4) is reminiscent of the firing patterns of grid cells. Both show a grid-like arrangement of regions with high activity. It is thus tempting to interpret SFA as a model for the self-organized formation of grid cells. Unfortunately, such an interpretation is superficial. Whereas grid cells show a hexagonal arrangement of firing fields, the grids of the SFA units are arranged in a rectangular fashion. In addition, the arrangements of the regions of high activity of the SFA units is strongly influenced by the boundary conditions and thus by the shape of the room. This is not the case for grid cells. Moreover, the spatial firing maps of grid cells show spatial frequencies between 39 and 73 cm (Hafting et al., 2005), whereas the spatial frequencies of the grids found in SFA depend on the size of the room and are unbounded in that solutions with high Δ-values can have arbitrarily high spatial frequencies. In summary, the response properties of the SFA units bear similarities to those of grid cell, but they lack a set of defining features of grid cells. It is possible though that a different implementation of the slowness principle and/or the constraints yields a better model (see chapter 8).

Previous studies have shown that place cells can be established by the unsupervised learning principle of sparse coding on the output of grid cells (Franzius, Vollgraf & Wiskott, 2007). Similarly, linear sparse coding can be used to learn units with similar response properties as place and head-direction cells from the output of the SFA units. Panel C in Figure 4.4 shows that sparse coding, applied to the SFA outputs for fast rotations leads to a localized representation of position that resembles place cells. The simulations in Figure 4.4 used cumulant-based independent component analysis (Blaschke & Wiskott, 2004) for sparse coding. Similar results can be obtained by means of biologically more plausible implementations of sparse coding, e.g., by competitive learning (Franzius, Sprekeler & Wiskott, 2007; Franzius, Vollgraf & Wiskott, 2007). Head-direction-like representations can also be learned by applying sparse coding to the SFA representations for slow rotation velocities. The results are shown in Panel F of Figure 4.4.

Interestingly, sparse coding on the SFA representation learned in a linear track generates a representation that shows both the localization and the direction-specificity of physiological place cells in linear tracks, as shown in panel C of Figure 4.5. Thus, the combination of slowness and sparseness can serve as a model of place cells and head-directions cells. In (Franzius, Sprekeler & Wiskott, 2007), we have also shown that the same model with different movement statistics reproduces the response properties of spatial view cells as found in primates (Rolls, 1999).

Although the representation of spatial coordinates as extracted by SFA are thus not immediately comparable with spatial representations found in the brain, they can be transformed into similar representations by rather simple transformations. The central achievement of SFA is the extraction of invariant spatial representations from complicated visual data. This problem is highly non-trivial and has been avoided by most previous sensory-driven models of place cells, which often use mode abstract input data. Thus, slowness learning is an interesting candidate mechanism for the self-organized formation of neural representations of space.

4.2.5 Effects of Inhomogeneous Movement Statistics

The optimal functions for both the open field and the linear track were calculated under the assumption that the probability distribution for position and velocity is spatially homogeneous. For real rats, this assumption will not be valid. Rats like to stay close to

walls, so the probability should be higher in these regions. The velocity distribution will in general not be homogeneous either. Close to the walls of the enclosure the velocity parallel to the wall tends to be higher than that perpendicular to the wall, since the opposite could lead to painful experiences. What is the effect of these inhomogeneities on the structure of the optimal solutions for SFA?

According to the perturbation theoretical treatment in section 3.3, the SFA solutions show spatial oscillations whose spatial frequency is determined by the variance distribution of the velocity. High velocities lead to lower spatial frequencies. A different interpretation would be that the spatial resolution of the solutions is higher in region and directions, where the velocity of the rat is small. Qualitatively, this property will be preserved by the sparse coding step. If we assume that the animal moves faster parallel to the wall of the arena than perpendicular to it, our theory thus predicts elongated place fields along the walls that might be similar to the crescent-shaped fields reported in (Muller, 1996) for a circular arena. The theory also predicts a reduction in the size of place fields associated with regions of the enclosure, where the animal moves more slowly, e.g., close to food sources. This prediction is difficult to test, however, because the animals tends to be at rest in these regions and the firing mode of place cells seems to change drastically when the animal is at rest (see e.g., Foster & Wilson, 2006).

4.2.6 Discussion

In this section, we have shown that SFA can extract high-level information such as the position and head direction of an animal from rather complicated visual data. The results for the open field indicate that the information encoded in the SFA output depends on the movement statistics of the rat. Fast translation leads to position-invariant units that encode the rat's head direction, while fast rotation leads to head-direction invariant units that encode the rat's position.

Using a linear sparse coding step, these representations can be transcoded into spatial representations that are similar to representations that were found in the hippocampal formation of rodents, i.e., place and head-direction cells. If the network learns place or head-direction cells depends on the movement statistics of the rat. For quick translation and slow head rotation, head-direction cells emerge, while slow translation with quick rotation yields place cells. But how can place cells and head-direction cells be learned simultaneously with just one movements statistics? In the article (Franzius, Sprekeler & Wiskott, 2007) we have proposed a solution to this question that is based on a learning rate adaptation of two different populations of cells. One population learns primarily during periods of relatively quick rotation while the other learns during periods of slow rotation. The first population develops place field characteristics while the latter resembles head-direction cells. The required modulation signal, essentially the ratio of translation and rotation velocity, could be provided by the vestibular system.

Further investigations will be needed to test if the extraction of position and head-direction is still reliable for more realistic stimuli, e.g., visual input recorded from a robot. In particular, natural visual scenes may contain additional parameters, such as changes in illumination, that vary more slowly than the animal's position or head direction. In this case, the output units of SFA would prefer to encode these features and become invariant to both head-direction and position of the animal. The output signals resulting from natural videos may be difficult to interpret, because the slowly changing features may not be known a priori.

The agreement of the theoretical predictions with the simulation results shows that the theory is valid not only for low-dimensional input data as for the nonlinear blind source separation problem in section 4.1, but also for rather high-dimensional visual input data lying on a low-dimensional manifold. This may be useful for other applications of SFA as well, e.g., for object recognition problems (Franzius, Wilbert & Wiskott, 2007). When using videos of a single rotating object for training, the output of SFA should represent the orientation of the object in space in an oscillatory fashion. The theory may thus help to extract a more convenient and compact representation of object orientation.

Chapter 5

Analytical Derivation of Complex Cell Properties

5.1 Introduction

About half a century has passed since the first characterization of the response behavior of cells in primary visual cortex (Hubel & Wiesel, 1962). Despite extensive research, both experimental and theoretical, the processes that shape the structure of their receptive fields are still a matter of debate. Several mechanisms for the self-organization of the early visual system have been proposed, ranging from genetically determined, 'hard-wired' (McLaughlin & O'Leary, 2005) or statistical connectivity patterns (Ringach, 2007) to optimal coding strategies claiming that receptive fields are learned from natural stimuli (e.g., Bell & Sejnowski, 1997; Olshausen & Field, 1996). Although both approaches can explain aspects of V1 receptive fields, they both suffer from a basic dilemma. On the one hand, the idea that receptive fields are learned from natural stimuli is challenged by experimental findings indicating that in some species receptive fields in V1 are largely developed when the animal first opens its eyes (Hubel & Wiesel, 1963)[1]. On the other hand, the notion that the early visual system is mostly hard-wired is problematic, because it has been shown that it remains plastic even in adulthood (Buonomano & Merzenich, 1998) and that receptive field properties can adapt to the statistics of artificial stimuli on short time scales down to minutes (Yao & Dan, 2001). Thus, V1 receptive fields must at least be compatible with natural stimuli because they would otherwise be unlearned quickly. One possible way of establishing this compatibility is that the prenatal development of the early visual system has adapted to natural stimuli on an evolutionary time scale. A different possibility is that spontaneous retinal activity occurring before eye-opening shapes the receptive fields (Wong, 1999; Cang et al., 2005; Torborg & Feller, 2005) and that there are intrinsic similarities between the statistics of retinal waves and natural stimuli. But what is the nature of these similarities and what type of learning rule could

[1]Note that the maturity of receptive fields shortly after birth may vary between cortical layers. Receptive fields of cells in layer IV, which are the main recipients of thalamocortical inputs, seem to be largely developed shortly after birth (Hubel & Wiesel, 1963). The receptive field maturity of cells in the superficial layers II/III is still debated. Albus & Wolf (1984) provided evidence that receptive fields of complex cells in layer II/III in cats are not fully developed until the 4th postnatal week, but this may not be true for primates, whose overall development is more advanced at birth. Moreover, there are indications that V1 cells can develop largely normal response properties in the absence of natural visual stimuli (Wiesel, 1982).

exploit them?

Most cells in V1 are activated by Gabor wavelets, i.e., by visual stimuli that resemble localized gratings. Based on the dependence of their firing rate on the phase of the grating and its contrast polarity, V1 cells are usually classified as *simple* or *complex cells*. Simple cells show a strong dependence on the phase of the stimulus while complex cells are largely independent with respect to stimulus phase. The phase invariance of complex cells can also be interpreted as an invariance to the position of the stimulus within the receptive field. Slow feature analysis, primarily designed for invariance learning, is thus a natural candidate for the self-organized formation of complex cell receptive fields.

A recent study has shown that slow feature analysis can indeed reproduce a wide range of properties of complex cell receptive fields (Berkes & Wiskott, 2005). For training, the authors used quasi-natural image sequences that were generated from static natural images by applying transformations such as translation, rotation, and zoom. The simulations yielded a set of quadratic functions that shared several properties with complex cells in V1 including grating-shaped optimal stimuli and different types of selectivity to orientation and frequency. What makes this study interesting in the context of the debate above is that the authors performed test experiments to evaluate which aspects of the training data were responsible for the structure of the simulated receptive fields. They found that higher order image statistics were immaterial (although they were accessible to the learning paradigm[2]) and that the same receptive fields could be learned with colored noise images. If the transformations that were used to generate the image sequences were changed however, the properties of the receptive fields changed. For example, Gabor-like optimal stimuli could not be obtained without translations in the training sequences. It is thus tempting to speculate that V1 receptive fields are not adapted to higher order statistics of natural stimuli but rather to transformations that typically occur in natural stimuli. Intriguingly, these transformations could also be present in retinal waves, as one could interpret propagating or rotating waves as an imitation of translations or rotations in natural stimuli.

In this chapter, we present a mathematical analysis of the simulations performed by Berkes and Wiskott and show that it is possible to derive several of the observed receptive field properties analytically. The analysis is based on the theory of Lie groups and, similar as in earlier chapters, leads to partial differential eigenvalue equations for the optimal functions of SFA. Since the framework is based on the transformations that were used to generate the image sequence and makes relatively weak assumptions about the image statistics, it further supports the idea that receptive fields could be shaped by image transformations rather than higher order statistics. Moreover, it gives an intuitive understanding for some of the response properties found in the simulations.

5.2 Theory

5.2.1 Assumptions & Notation

The Function Space

We consider the case where the input data are continuous grey-scale images, with $x(\mathbf{r})$ denoting the grey value at pixel position $\mathbf{r}$. This implies that the input data are infinite-

[2]SFA with quadratic functions is based on correlations of polynomials of degree two in the images, and can thus access fourth order statistics.

dimensional, so the input-output functions g_j for SFA are functionals that map images to real numbers. Berkes & Wiskott (2005) used quadratic functions, i.e., sums of monomials of first and second order in the pixel values. We will later focus on functions that are translation-invariant. The only linear function that is translation-invariant is the mean pixel intensity, which is not very informative about the image. Therefore, we neglect the linear component and restrict the function space $\mathcal{F}$ to the space of quadratic functionals of the images:

$$g[x(\mathbf{r})] = \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} g(\mathbf{r}, \mathbf{r}') x(\mathbf{r}) x(\mathbf{r}') \mathrm{d}^2 r \, \mathrm{d}^2 r' \,, \tag{5.1}$$

where $g(\mathbf{r}, \mathbf{r}') = g(\mathbf{r}', \mathbf{r})$ is a symmetric function. For mathematical convenience, we assume that the images are infinitely large, so the integrals extend over $\mathbb{R}^2$. In the following we will refer to the elements of $\mathcal{F}$ as functions, although they would more commonly be referred to as functionals. Note that $g(\mathbf{r}, \mathbf{r}')$ can be understood as the representation of the functional $g[x(\mathbf{r})]$ in terms of the basis functionals $x(\mathbf{r})x(\mathbf{r}')$.

Note that due to the restriction of the function space $\mathcal{F}$ and the infinite dimension of the training data, the theory developed in chapter 3 cannot describe this applications of SFA.

Input Data Generation: Transformation Group

Berkes & Wiskott (2005) generated their training data from natural static images by shifting, rotating, and zooming a quadratic frame across the images, thus generating a set of image sequences that display transformations that are typical for natural image sequences. We will use the same paradigm here. Because the images are assumed to be infinitely large, all of these transformations are invertible. They are also continuous and smooth, as e.g., translations can be made with arbitrarily small shifts. Mathematically, this implies that the transformations form a Lie group, i.e., a continuous group. The generation of one image sequence of the training data can be written as

$$x(\mathbf{r}, t) = T_x(t) x(\mathbf{r}, 0) \,, \tag{5.2}$$

where $T_x(t)$ is an operator that maps the image at time $t = 0$ to the image at time t. The set of all possible operators T_x forms a representation of the transformation group on the vector space of the images. The effect of the transformation operators for translation, rotation, and zoom on an image x are given in table 5.1.

transformation	**effect**
translation by a vector $\mathbf{R}$	$\left(T_x^{\mathbf{R}} x\right)(\mathbf{r}) = x(\mathbf{r} - \mathbf{R})$
rotation with an orthogonal matrix $\mathbf{O}$	$\left(T_x^{\mathbf{O}} x\right)(\mathbf{r}) = x(\mathbf{O}^{-1}\mathbf{r})$
zoom by a factor z	$\left(T_x^z x\right)(\mathbf{r}) = x(\mathbf{r}/z)$

Table 5.1: Effect of the transformations operators for translation, rotation, and zoom on an image $x(\mathbf{r})$.

A different representation of the transformation group can be constructed by defining operators T_g that act on the functions in $\mathcal{F}$ such that

$$(T_g g)[x(\mathbf{r})] := g[T_x x(\mathbf{r})] \tag{5.3}$$

is fulfilled for all functions $g \in \mathcal{F}$ and all images $x(\mathbf{r})$. Intuitively, this representational change corresponds to a change of the coordinate system. Let us think of the function g as a measurement device that extracts certain aspects of the image. Then instead of moving the image the function g acts on (this is the effect of T_x) one may also move the function in the opposite direction (this is the effect of T_g). In the following we will skip the subscript g, as all operators will act on functions, not on images.

The Hilbert Space of Functions

The function space $\mathcal{F}$ obviously forms a vector space. It is convenient to turn it into a Hilbert space by defining the scalar product

$$(f, g) = \langle f[x(\mathbf{r}, t)] g[x(\mathbf{r}, t)] \rangle \,, \tag{5.4}$$

where the average $\langle \cdot \rangle$ is taken over the training data, i.e., over all sequences and times within the sequences. For simplicity of notation, we will in the following omit the average over sequences and act as if there was only one sequence. All our considerations are valid for an ensemble of sequences as well, but many quantities would need additional indices that would only cause confusion. For the same reason, we will often skip the argument of the functions g.

Note that the scalar product (5.4) is the same as the scalar product (3.13) defined in chapter 3. Thus, if the functions have zero mean on the training data this scalar product measures the covariance between the output of the functions f and g. Consequently, the unit variance and decorrelation constraints (2.3, 2.4) again take the compact form of an orthonormality constraint

$$(g_i, g_j) = \delta_{ij} \,, \tag{5.5}$$

where δ_{ij} denotes the Kronecker symbol.

For all the derivations that follow, it is assumed that the scalar product exists (i.e., that it is finite) for all functions f and g it acts upon. This excludes, e.g., functions with infinite variance and thus inflicts constraints on the function space $\mathcal{F}$. Note that these constraints may also depend on the space of images from which the image sequences are generated. Together with equation (5.1), the constraints define the function space on which the scalar product can be defined.

Invariance of the Training Data

In the following we will assume that the statistics of the input data are invariant with respect to the transformations applied. There are two arguments why this assumption is reasonable. Firstly, natural image statistics seem to be largely translation and rotation invariant and show some degree of scale invariance (Ruderman & Bialek, 1994; Dong & Atick, 1995; Dong, 2001)[3]. Secondly, the training data are generated by applying these transformations, so there is no reason why a transformed version of the image should be less likely than the original one.

[3]There is a weak dependence of natural image statistics on orientation, however. In accordance with the scale invariance of natural images, the power spectrum resembles a power law, the associated exponent of which depends on orientation (Ruderman & Bialek, 1994). The underlying cause is the frequent occurrence of vertical and horizontal edges in natural scenes (e.g., trunks of trees, the horizon). In these directions, the correlation functions decay more slowly.

The invariance of the image statistics means that if the whole ensemble of images used for training is subjected to any of the transformations, the resulting ensemble of images has the exactly the same statistics. In particular, the moments of the image statistics remain the same. Then, the scalar product between two functions f and g has to take the same value for the transformed ensemble as well, because it is merely an average of a nonlinear function of the images and thus a linear combination of the moments of the image statistics. Mathematically, this implies that the scalar product (5.4) is invariant with respect to all operators T in the transformation group:

$$(Tf, Tg) = (f, g)\,. \tag{5.6}$$

In other words, the transformation operators T are orthogonal with respect to the scalar product (5.4), i.e., they preserve distances (i.e., the standard deviation of differences of output signals) and angles (which are related to the correlation of output signals) in the function space $\mathcal{F}$ as derived from the scalar product (5.4).

Generators of the Transformations

The transformation operators T form a manifold, embedded in the space of all linear operators on the function space $\mathcal{F}$. This manifold is of relatively low dimension. For example, if we use translation, rotation and zoom, the operators can be characterized by the translation vector (two degrees of freedom), the rotation angle, and the zoom factor. The operator manifold would thus be 4-dimensional. Nevertheless, the manifold can in principle have a very complicated structure, so that its low dimensionality is not necessarily helpful. Here, we will show that the low dimensionality of the manifold in combination the invariance assumption introduced above has important implications for the temporal derivative of time-dependent transformations. We will introduce the rather mathematical concept of *generators* of transformations, which will be useful for a reformulation of the slowness objective in the next section.

SFA focusses on the temporal derivative of the output signals. In the transformation operator notation, the output signals within one image sequence are generated by the prescription

$$y(t) = (T(t)g)[x(\mathbf{r}, 0)]\,. \tag{5.7}$$

Taking the derivative yields

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = \left(\frac{\mathrm{d}}{\mathrm{d}t}T(t)g\right)[x(\mathbf{r}, 0)] \tag{5.8}$$

$$=: (T(t)Q(t)g)[x(\mathbf{r}, 0)]\,, \tag{5.9}$$

with $Q(t) := T^{-1}(t)\left[\frac{\mathrm{d}}{\mathrm{d}t}T(t)\right]$.

The set of possible operators $Q(t)$ has two properties that will be useful in the following. First, it can be shown that $Q(t)$ is an element of the tangent space of the transformation group at the identity element:

Theorem 3. *Let $T(t)$ be a differentiable trajectory of transformations with $T(t)$ element of a Lie transformation group for all t. Then for all t, $Q(t) = T^{-1}(t)\left[\frac{d}{dt}T(t)\right]$ is an element of the tangent space of the transformation group at the identity element.*

Proof. It is sufficient to show that there is a trajectory $\tilde{T}(s)$ of transformation operators such that $\tilde{T}(s)|_{s=t} = E$ and $\frac{\mathrm{d}}{\mathrm{d}s}\tilde{T}(s)|_{s=t} = Q(t)$. It is easy to see that $\tilde{T}(s) := T^{-1}(t)T(s)$ fulfills these conditions. □

This implies that all $Q(t)$ are elements of a vector space that has the same (low) dimensionality as the transformation group. We can thus choose a basis G_α of this vector space and write

$$Q(t) = \sum_\alpha v_\alpha(t) G_\alpha \,. \tag{5.10}$$

The basis operators G_α are often referred to as the *generators* of the transformation group. For reasons that will become obvious later, we will refer to the coefficients v_α as *velocities*. The generators G_α will play a central role in the derivation of the optimal functions.

Second, the fact that the operators T are orthogonal (because the image statistics are invariant under the transformations) implies that the generators G_α are anti-selfadjoint:

Theorem 4. *The generators G_α are anti-selfadjoint with respect to the scalar product (5.4), i.e.,*

$$(f, G_\alpha g) = -(G_\alpha f, g) \tag{5.11}$$

for all $f, g \in \mathcal{F}$.

Proof. G_α is an element of the tangent space of the transformation group at the identity element. Thus, there is a trajectory $T(s)$ of transformation operators such that $\left[\frac{\mathrm{d}}{\mathrm{d}s}T(s)\right]_{s=0} = G_\alpha$ and $T(0) = E$. Because $T(s)$ is orthogonal for all values of s, $(T(s)f, T(s)g) = (f, g)$ is independent of s for arbitrary f, g. Thus

$$\begin{aligned} 0 &= \left[\frac{\mathrm{d}}{\mathrm{d}s}(T(s)f, T(s)g)\right]_{s=0} && (5.12)\\ &= \left[\left(\frac{\mathrm{d}}{\mathrm{d}s}T(s)f, T(s)g\right) + \left(T(s)f, \frac{\mathrm{d}}{\mathrm{d}s}T(s)g\right)\right]_{s=0} && (5.13)\\ &= (G_\alpha f, g) + (f, G_\alpha g)\,, && (5.14) \end{aligned}$$

which proves the assertion. □

5.2.2 Reformulation of the Slowness Objective

The conventions introduced in the last section allow us to rewrite the slowness objective (2.1):

$$\begin{array}{lcl} \Delta(g) & := & \langle \dot{y}(t)^2 \rangle \\ & \overset{(5.9,5.10)}{=} & \sum_{\alpha,\beta} \langle v_\alpha(t)(T(t)G_\alpha g)[x(\mathbf{r},0)] v_\beta(t)(T(t)G_\beta g)[x(\mathbf{r},0)] \rangle \,. \end{array} \tag{5.15}$$

Assuming that the velocities v_α are statistically independent of the transformation, we can split the average and express the Δ-value in the form of a scalar product:

$$\begin{array}{lcl} \Delta(g) & \overset{(5.15)}{=} & \sum_{\alpha,\beta} \langle v_\alpha(t) v_\beta(t) \rangle \langle (T(t)G_\alpha g)[x(\mathbf{r},0)] (T(t)G_\beta g)[x(\mathbf{r},0)] \rangle \\ & \overset{(5.3,5.2)}{=} & \sum_{\alpha,\beta} \langle v_\alpha(t) v_\beta(t) \rangle \langle (G_\alpha g)[x(\mathbf{r},t)] (G_\beta g)[x(\mathbf{r},t)] \rangle \end{array} \tag{5.16}$$

$$\begin{aligned} &\overset{(5.4)}{=} && \sum_{\alpha,\beta}\langle v_\alpha(t)v_\beta(t)\rangle(G_\alpha g, G_\beta g) \\ &\overset{(5.11)}{=} && \Big(g, \underbrace{\Big[-\sum_{\alpha,\beta}\langle v_\alpha v_\beta\rangle G_\alpha G_\beta\Big]}_{=:\mathcal{D}} g\Big) \\ &= && (g, \mathcal{D}g)\,. \end{aligned} \tag{5.17}$$

Because the operator $\mathcal{D}$ is a bilinear combination of the anti-selfadjoint generators G_α, it itself is self-adjoint, i.e.

$$(f, \mathcal{D}g) = (\mathcal{D}f, g) \quad \forall f, g \in \mathcal{F}\,. \tag{5.18}$$

5.2.3 A Differential Equation for the Optimal Solutions

The main advantage of the reformulation of the objective function is that the optimization problem that underlies SFA takes a very convenient form. Just as in chapter 3, the functions that minimize (5.17) are the eigenfunctions of the operator $\mathcal{D}$, while the fact that $\mathcal{D}$ is self-adjoint ensures that the eigenfunctions are orthonormal so that the constraints (5.5) are fulfilled (for the relevant mathematical background see e.g. (Landau & Lifshitz, 1977, §20) or (Courant & Hilbert, 1989)). The eigenvalues Δ_j are the Δ-values of the eigenfunctions. We can thus solve the optimization problem of SFA by finding the J solutions of the eigenvalue equation

$$\mathcal{D}g_j = \Delta_j g_j \tag{5.19}$$

with the smallest eigenvalues Δ_j.

The first important result of this chapter is that this equation does not depend on higher order statistics of the images used for training. Rather, it depends on the nature of the transformations underlying the image sequences (as reflected by the generators G_α) and the second order moments $\langle v_\alpha v_\beta\rangle$ of the associated velocities v_α. This explains the finding by Berkes & Wiskott (2005) that the simulated receptive fields for training sequences that were generated from colored noise images and those for sequences generated from natural images were essentially the same. On the other hand, a change in the transformations used to generate the image sequences changes the structure of the operator $\mathcal{D}$ and thus the resulting receptive fields. This is also in agreement with the simulations. It is important to bear in mind that equation (5.19) is only valid if the input statistics are invariant with respect to the transformations. If this invariance condition is not fulfilled, higher order statistics may play a role.

Solving (5.19) of course requires that we know $\mathcal{D}$, which according to equation (5.17) corresponds to knowing the generators G_α of the transformations and the matrix of the second moments $\langle v_\alpha v_\beta\rangle$ of the associated velocities v_α. For reasons of compactness, we defer the derivation of the generators to appendix A.2. Suffice to say that we represent G_α such that they act on the kernel $g(\mathbf{r}, \mathbf{r}')$ of the quadratic functionals (5.1). In this representation they become the differential operators listed in table 5.2. The associated velocities are the translation velocity $\mathbf{v}$, the angular velocity ω for rotation, and a zoom velocity ζ for zoom. Note that we interpret the zoom velocity as the factor by which the size of the image increases per time unit. Let z denote the factor by which an image has been zoomed relative to its original size. Then constant zoom velocity implies that the zoom factor z grows exponentially in time, so that not $\dot{z}$ is constant, but rather $\frac{\dot{z}}{z} =: \zeta$.

transformation	**generator**	**velocity**
translation	$\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'}$	$\mathbf{v}$
rotation	$r_1\partial_{r_2} - r_2\partial_{r_1} + r'_1\partial_{r'_2} - r'_2\partial_{r'_1}$	ω
zoom	$\nabla_{\mathbf{r}} \cdot \mathbf{r} + \nabla_{\mathbf{r}'} \cdot \mathbf{r}' = \mathbf{r} \cdot \nabla_{\mathbf{r}} + \mathbf{r}' \cdot \nabla_{\mathbf{r}'} + 4$	$\zeta = \dot{z}/z$

Table 5.2: Generators of the transformations used to generate the image sequences. ∂_{r_1} denotes the derivative with respect to the first component of $\mathbf{r}$. $\nabla_{\mathbf{r}}$ denotes the vector-valued operator $(\partial_{r_1}, \partial_{r_2})^T$.

With these generators the eigenvalue equation becomes a partial differential eigenvalue equation for $g(\mathbf{r}, \mathbf{r}')$. Finding a closed form general solution is rather difficult, mainly because the resulting image depends on the order in which translation and rotation/zoom are applied. A mathematical implication is that the generators for translation and those for rotation and zoom do not commute, so they do not possess a common set of eigenfunctions (which would simplify the analysis significantly). However, in the special case of translation-invariant functions, it is possible to find a closed form solution that explains the orientation and frequency dependence of the simulated receptive fields in (Berkes & Wiskott, 2005) as well as aspects of their optimal stimuli.

5.3 Results

5.3.1 Translation-Invariant Solutions

In this section, we will derive and discuss an explicit solution of the eigenvalue problem (5.19) for the special case of translation-invariant functions. But why translation invariance?

The control experiments performed in (Berkes & Wiskott, 2005) suggest that translation is a necessary and sufficient condition for the optimal functions to resemble complex cells. In simulations where translation was present in the training data, the functions became phase-invariant and had optimal stimuli that resemble Gabor wavelets. The invariance of the units to spatial phase corresponds to a certain degree of translation invariance. Because of this and because the eigenvalue equation (5.19) can be solved analytically for this case, we will apply the theory to the special case where the functions g are translation invariant.

Mathematically, translation-invariance implies that the functions $g(\mathbf{r}, \mathbf{r}')$ depend on the difference $\mathbf{r} - \mathbf{r}'$ only: $g(\mathbf{r}, \mathbf{r}') = \tilde{g}(\mathbf{r} - \mathbf{r}')$. In this case, the output signal depends on the power spectrum of the image only, because

$$g[x(\mathbf{r})] = \int \tilde{g}(\mathbf{r} - \mathbf{r}')x(\mathbf{r})x(\mathbf{r}')\mathrm{d}^2r\mathrm{d}^2r' \tag{5.20}$$

$$= \int \tilde{g}(\mathbf{k})|x(\mathbf{k})|^2\mathrm{d}^2k\,. \tag{5.21}$$

Here $x(\mathbf{k}) := \frac{1}{2\pi}\int x(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}\mathrm{d}^2r$ and $\tilde{g}(\mathbf{k})$ denote the Fourier transforms of the image and the function $\tilde{g}(\mathbf{r})$, respectively. The value $|x(\mathbf{k})|^2$ of the power spectrum of an image x is calculated by summing the squares of the sin-Fourier transform and the cos-Fourier transform. Therefore, equation (5.21) implies that the function g is a weighted sum of quadrature filter pairs with filters that are place waves. Note that quadrature filter pairs are the key element of the standard "energy" model of complex cells (Adelsen & Bergen, 1985).

Another implication of the translation-invariance of g is that it is an eigenfunction of the generator of translations with the eigenvalue 0:

$$(\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'})g(\mathbf{r}, \mathbf{r}') = (\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'})\tilde{g}(\mathbf{r} - \mathbf{r}') = 0\,. \tag{5.22}$$

We can thus neglect the contribution of the translation generator in the eigenvalue equation (5.19).

In the simulations, the transformation velocities (i.e., the differences in position, angle, and zoom factor between successive frames) were chosen independently and from Gaussian distributions centered at zero. The matrix $\langle v_\alpha v_\beta \rangle$ is then diagonal and contains the mean squares of the velocities on the diagonal. Neglecting terms arising from translation, the eigenvalue equation (5.19) then takes the form:

$$-\left[\langle\omega^2\rangle (G^{\text{rot}})^2 + \langle\zeta^2\rangle (G^{\text{zoom}})^2\right] g_j = \Delta_j\, g_j\,. \tag{5.23}$$

Because g depends only on $\mathbf{r} - \mathbf{r}'$, it is convenient to use a center-of-mass coordinate system by defining $\mathbf{R} := \frac{1}{2}(\mathbf{r}+\mathbf{r}')$ and $\tilde{\mathbf{r}} := \mathbf{r}-\mathbf{r}'$. In this coordinate system the generator for translation becomes $G_{\text{trans}} = \nabla_{\mathbf{R}}$, so that equation (5.22) takes the form

$$\nabla_{\mathbf{R}}\, g(\mathbf{R} + \tilde{\mathbf{r}}/2, \mathbf{R} - \tilde{\mathbf{r}}/2) = \nabla_{\mathbf{R}}\tilde{g}(\tilde{\mathbf{r}}) = 0\,, \tag{5.24}$$

i.e., g is independent of $\mathbf{R}$. The generators for rotation and zoom become

$$G^{\text{rot}} = R_1\partial_{R_2} - R_2\partial_{R_1} + \tilde{r}_1\partial_{\tilde{r}_2} - \tilde{r}_2\partial_{\tilde{r}_1} \tag{5.25}$$

$$G^{\text{zoom}} = \mathbf{R}\cdot\nabla_{\mathbf{R}} + \tilde{\mathbf{r}}\cdot\nabla_{\tilde{\mathbf{r}}} + 4\,. \tag{5.26}$$

For translation-invariant functions the components of the generators that contain derivatives with respect to $\mathbf{R}$ can be neglected, which leads to a further simplification of equation (5.23). The solution for the resulting eigenvalue equation can be given in a closed form. Because the behavior of the functions g is much easier to discuss in the Fourier representation (5.21), however, it is more convenient to solve the eigenvalue equation for the Fourier transform $\tilde{g}_j(\mathbf{k})$ directly. Transferring the eigenvalue equation into Fourier space requires a long, but not very illustrative derivation. Essentially, we have to insert the generators (5.25) and (5.26) and the definition of the Fourier transform of $\tilde{g}$ into equation (5.23) and use the property of the Fourier transform that multiplications with $\tilde{\mathbf{r}}$ correspond to derivatives with respect to $\mathbf{k}$ in Fourier space and that derivatives with respect to $\tilde{\mathbf{r}}$ become multiplications with $\mathbf{k}$. For brevity, we skip the details and simply state the resulting eigenvalue equation:

$$-\left[\langle\omega^2\rangle (k_1\partial_{k_2} - k_2\partial_{k_1})^2 + \langle\zeta^2\rangle(\mathbf{k}\cdot\nabla_{\mathbf{k}} - 2)^2\right]\tilde{g}_j(\mathbf{k}) = \lambda_j\tilde{g}_j(\mathbf{k})\,. \tag{5.27}$$

It is easier to solve this equation in polar coordinates $(k, \phi) \in \mathbb{R}^+ \times [0, 2\pi[$, because then the operators for translation and rotation separate:

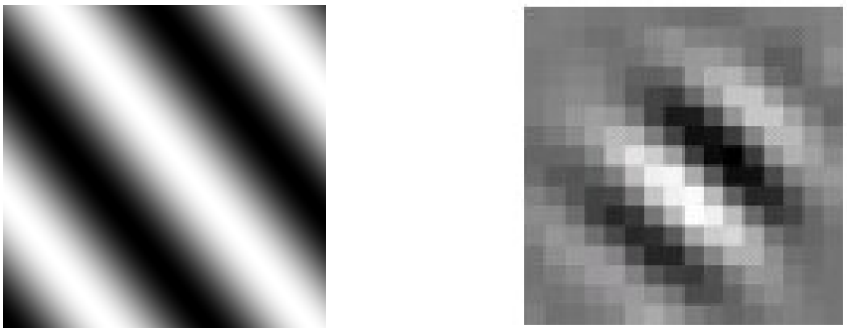
$$-\left[\langle\omega^2\rangle\partial_\phi^2 + \langle\zeta^2\rangle(k\partial_k - 2)^2\right]\tilde{g}_j(k,\phi) = \Delta_j\tilde{g}_j(k,\phi)\,. \tag{5.28}$$

The eigenfunctions to this equation are given by

$$\tilde{g}_{q,m}(k,\phi) = A_{q,m}k^2 Q_q(k) M_m(\phi) \tag{5.29}$$

$$\text{with}\quad Q_q(k) = \begin{cases} \cos(q\ln k) & \text{for } q \geq 0 \\ \sin(q\ln k) & \text{for } q < 0 \end{cases} \tag{5.30}$$

$$\text{and}\quad M_m(\phi) = \begin{cases} \cos(m\phi) & \text{for } m \text{ even} \\ \sin((m+1)\phi) & \text{for } m \text{ odd} \end{cases}, \tag{5.31}$$



(a) Theory (b) Simulation

Figure 5.1: **Optimal stimuli.** (a) The theoretically predicted optimal stimuli are delocalized plane waves. (b) Typical optimal stimulus for the simulated SFA units in (Berkes & Wiskott, 2005). As theoretically predicted, the unit responds most strongly to a grating with a specific orientation. The observed decay of the simulated optimal stimulus towards to boundary of the image patch, however, is not captured by the theory.

and the associated eigenvalues are given by

$$\Delta_{q,m} = \langle \zeta^2 \rangle q^2 + \begin{cases} \langle \omega^2 \rangle m^2 & \text{for } m \text{ even} \\ \langle \omega^2 \rangle (m+1)^2 & \text{for } m \text{ odd} \end{cases} . \tag{5.32}$$

$A_{q,m}$ denotes a normalization constant that ensures that the unit variance constraint is fulfilled for the training data at hand and $q \in \mathbb{R}$ and $m \in \mathbb{N}^0$ are indices that label the solution. Note that the oscillation in the angular direction contains only even frequencies (m for m even and $m+1$ for m odd), because $\tilde{g}(\tilde{\mathbf{r}})$ is real-valued and symmetric, so its Fourier transform has to be symmetric, i.e., $\tilde{g}(k,\phi) = \tilde{g}(k,\phi+\pi)$. For each Δ-value, there are four solutions, corresponding to all possible combinations of sine and cosine in k and ϕ.

Note that in addition to those given in equation (5.29) there are also solutions that have negative eigenvalues. These solutions have a frequency dependence that follows $\tilde{g} \approx k^2 e^{q \ln(k)} = k^{2+q}$ with $q \in \mathbb{R}$. These functions cannot be normalized because they have infinite variance. We thus exclude them as possible solutions.

5.3.2 Optimal Stimuli

We define the optimal excitatory stimulus of a function $g[x(\mathbf{r})]$ as the image $S^+(\mathbf{r})$ that maximizes $g[x(\mathbf{r})]$ under the constraint of fixed total image power

$$\int S^+(\mathbf{r})^2 \mathrm{d}^2 r = \int |S^+(\mathbf{k})|^2 \mathrm{d}^2 k = \text{const.} \tag{5.33}$$

Similarly, the optimal inhibitory stimulus $S^-(\mathbf{r})$ is the image that minimizes $g[x(\mathbf{r})]$ with fixed power. According to (5.21), translation-invariant quadratic functionals are linear functionals of the power spectrum $|x(\mathbf{k})|^2$, so it is intuitively clear that the optimal excitatory/inhibitory stimulus concentrates all its power to those frequencies where $\tilde{g}_{q,m}(\mathbf{k})$ is maximal/minimal. This has several implications:

(a) The optimal excitatory/inhibitory stimuli are (possibly linear combinations of) plane waves $S^{\pm}(\mathbf{r}) = \cos(\mathbf{k} \cdot \mathbf{r} + \text{phase shift})$ with wave vectors $\mathbf{k}$ for which $\tilde{g}_{q,m}(\mathbf{k})$

is maximal/minimal. In practice, $\mathbf{k}$ will be restricted to a finite domain, in particular because the finite resolution introduces a frequency cutoff. $\tilde{g}_{q,m}$ will have at least one maximum within this domain. For large m, $\tilde{g}_{q,m}$ has many maxima of equal value, but in practice, one of these maxima will be slightly higher, so that the optimal stimulus will be a single plane wave. This agrees with the observations by Berkes & Wiskott (2005).

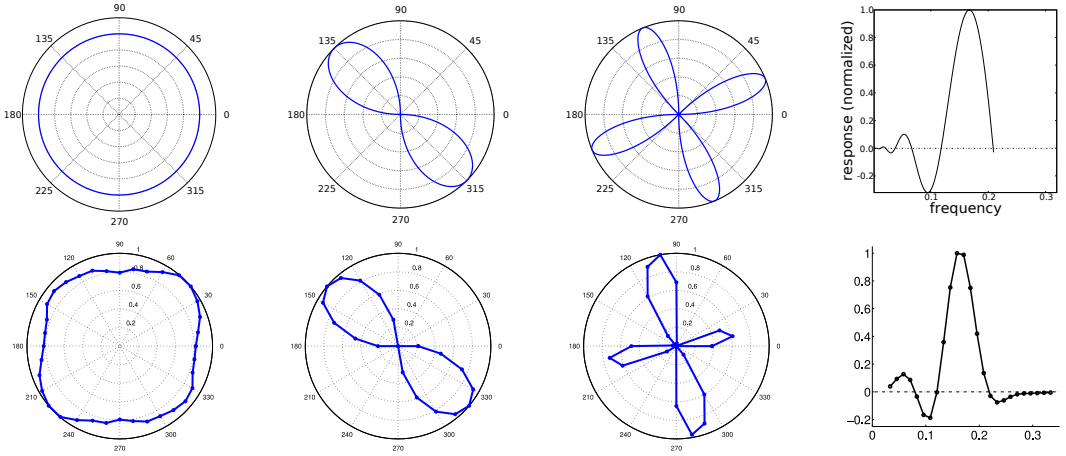
(b) The phase of the plane waves is arbitrary, because the functions g depend only on the power spectrum of the images and not on its phase structure. This is in line with the notion of complex cells as being invariant with respect to the phase of their optimal stimulus and is also consistent with the results by (Berkes & Wiskott, 2005).

(c) Since all functions $\tilde{g}_{q,m}(\mathbf{k})$ rise quadratically with the frequency $k = |\mathbf{k}|$, high frequencies are favored. In experiments with real data, there will of course be a frequency cutoff due to the finite resolution of the images, so that the optimal stimuli cannot have arbitrarily high frequencies. Berkes & Wiskott (2005) used principal component analysis (PCA) to reduce the input dimensionality from two pictures with 16×16 pixels each to a total of 100 dimensions, i.e., approximately 50 dimensions per image. It is known that PCA on natural images concentrates on low spatial frequencies while neglecting high frequencies. The highest spatial frequency that is possible after this preprocessing is then on the order of $\sqrt{50}/2 \approx 3.5$ cycles per side length of the image patch. This is a rather accurate estimate of the frequency of the optimal stimuli that were found by Berkes & Wiskott (2005).

(d) Unlike in physiological findings, the optimal stimuli are not localized. Rather intuitively, this can be understood as follows: In the case of image sequences that are generated by continuous transformations, i.e., where image content stays within the vicinity of its original position for a certain time, spatial integration effectively acts as a low-pass filter, with spatial integration over larger areas corresponding to low-pass filtering with longer time scales. Because low-pass filtering with longer time scales will generally lead to slower signals, optimal functions for SFA will always try to integrate over the largest area possible, i.e., the full image. This is reflected by the delocalized optimal stimuli. Note that SFA does not allow low-pass filtering, but requires the functions to process the input instantaneously. The low-pass filtering discussed here is purely spatial in nature but has the same effect as a temporal low-pass filter due to the spatiotemporal correlation structure of the input signals. Note also that the apparent localization of the simulated optimal stimuli is not a real localization as found, e.g., by (Olshausen & Field, 1996). The optimal functions decay towards the boundary of the image patch in order to reduce the influence of new pixels that abruptly enter the image patch. The optimal stimuli should thus vanish on the boundary as well. The optimal stimuli found by Berkes & Wiskott (2005) are as delocalized as this constraint allows them to be.

5.3.3 Orientation and Frequency Tuning

The typical approach for testing the orientation and the frequency tuning of a cortical cell is to plot its response to a grating as a function of the orientation and the frequency of the grating (see, e.g, De Valois et al., 1982). We represent the grating by a plane



(a) non-oriented units (b) orthogonal inhibition (c) secondary response lobes (d) frequency tuning

Figure 5.2: **Orientation and frequency tuning.** Upper row: Analytical results. Lower row: Simulation results (with permission from Berkes & Wiskott (2005)). **(a)**, **(b)** and **(c)**: Orientation tuning of the analytical solutions $\tilde{g}_{q,m}$ for $m = 0, 2$ and 4, compared with the orientation tuning of three units as simulated by Berkes & Wiskott (2005). Negative responses were truncated. Cells with similar orientation tuning were also observed in primary visual cortex of the macaque (De Valois et al., 1982), for a comparison of experimental and simulation results see (Berkes & Wiskott, 2005). The deviations of the simulation results from the theoretical preditions are due to random correlations in the input data that lead to a weak mixing of the theoretical solutions. For example, the simulated orientation tuning in (a) is a combination of the theoretical predictions in (a) ($m = 0$) and (c) ($m = 4$). The amplitude difference of the lobes of the simulation results in (c) can be explained by a mixture of the theoretical solutions in (b) ($m = 2$) and (c)($m = 4$). **(d)** Frequency tuning. For the theoretical results in the upper panel, the parameter q, the phase of the oscillation and the cutoff frequency were adapted to match the simulation results.

wave with frequency k_0 and orientation ϕ_0. As the power spectrum of this function is δ-shaped, the output of the function $g_{q,\phi}[x]$ for a plane wave is given by $\tilde{g}_{q,m}(k_0, \phi_0)$.

Figure 5.2 shows a comparison of the orientation and frequency tuning of the analytical solutions and the simulations. They are in good agreement apart from a frequency cutoff in the simulations that arises from the finite resolution of the images and the preprocessing (see discussion in section 5.3.2). The fact that the analytical solutions agree with the simulations indicates that the orientation and frequency tuning as observed in the simulations is an effect of the transformations used to generate the image sequences. But is there an intuitive understanding why the tuning curves have this shape?

The key to this question is the earlier result by Wiskott (2003) that the optimal output signals for SFA are harmonic oscillations (see also section 3.4). It is obvious that the output signal of the functions $\tilde{g}_{q,m}$ when applied to an image that rotates with constant velocity is sinusoidal. Similarly, the frequency dependence is such that the output signal is sinusoidal if the image is subjected to zoom with constant velocity. Remember that constant zoom velocity ζ implies that the zoom factor $z(t) = \exp(\zeta t)$ increases exponentially. As the image is zoomed by a factor z, the frequency decreases as $1/z$, so with an exponentially increasing zoom factor, the frequencies also decrease exponentially. In combination with the logarithmic dependence of $Q_q(k)$ on the frequency k, this yields a harmonic oscillation.

The reason for the quadratic rise of the oscillation amplitude of $\tilde{g}_{q,m}$ as a function of

the frequency k is more subtle. When the image is zoomed by a factor z (i.e., $x(\mathbf{r}) \to x(\mathbf{r}/z)$), the total image power P increases by a factor z^2. This can be seen by means of a coordinate transformation $\mathbf{r}' = \mathbf{r}/z$:

$$\text{power zoomed} = \int |x(\mathbf{r}/z)|^2 \mathrm{d}^2 r = \int |x(\mathbf{r}')|^2 z^2 \mathrm{d}^2 r' = z^2 \times \text{power unzoomed}\,. \quad (5.34)$$

The additional factor k^2 counterbalances the increase in the output signal that would normally result from the increase in power, so that the amplitude of the harmonic oscillation remains constant.

5.4 Discussion

In this section, we have presented a mathematical analysis of the simulations of complex cell receptive fields that were presented by Berkes & Wiskott (2005). The theory is based on a group-theoretical approach that focusses on the transformations that are typically present in natural visual scenes. It culminates in a partial differential eigenvalue problem that could be solved analytically for the special case of translation-invariant receptive fields. The orientation and frequency tuning of the analytical solutions are in good agreement with the simulation results. Moreover, the optimal stimuli of the analytical solutions are plane waves, similar to the gratings that were found in the simulations and that are also common in physiological studies of cells in V1.

Under the assumption that the statistics of the input image sequences are invariant with respect to the transformations used for their generation, the equations that determine the optimal functions are independent of the input statistics. Instead, they depend solely on the transformations as reflected by their group-theoretical generators. This purely mathematical statement agrees with control experiments performed by Berkes & Wiskott (2005), which showed that the simulation results were qualitatively the same when using colored noise instead of natural images. Which transformations were used, however, had a drastic influence on the structure of the receptive fields. For example, a lack of translation abolished the grating structure of the optimal stimuli. This is in agreement with the theory, because the optimal stimuli were plane waves only because the functions were assumed to be translation invariant. The assumption of translation invariance, however, is only valid when translation is the dominant transformation in the image sequences, so that any dependence on position would yield quickly varying output signals and would thus be unfavorable for the slowness objective.

The theory shows that each of the properties of the optimal functions can be understood as an effect of one particular transformation: Translation leads to optimal stimuli that are plave waves, rotation causes a sinusoidal dependence of the output on the orientation, and zoom is responsible for the frequency tuning of the cells. Intuitively, both the orientation and the frequency tuning can be understood as a way to generate harmonically oscillating output signals when the associated transformation is applied with constant velocity. This interpretation is in line with earlier results indicating that the optimal output signals for SFA are harmonics oscillations (Wiskott, 2003).

One property of complex cells in visual cortex is captured neither by the simulations nor by the theory: Receptive fields of cells in primary visual cortex are localized. As discussed in section 5.3.2, however, this cannot be expected from the slowness principle alone, because larger receptive fields allow slower responses, so localization is not favorable from the perspective of the slowness principle. Localized receptive fields probably

require either a different implementation of the constraints or additional objectives such as sparseness or statistical independence, which have been proposed as principles for the unsupervised learning of localized receptive fields of simple cells in V1 (Bell & Sejnowski, 1995; Olshausen & Field, 1996, 1997).

The optimal stimuli found in the simulation seem to possess at least some kind of localization, since they decay towards to borders of the images patch. A similar decay of the optimal functions towards the boundaries was also observed in a simpler SFA-based model of visual processing (Wiskott & Sejnowski, 2002). These results suggest that for the case where the input images are not infinitely large, the differential equation for the optimal functions has to be complemented by a boundary condition that requires the kernel of the optimal functionals to vanish on the boundary. Such a boundary condition would weaken the effect of new image pixels that enter the receptive field at its border and thus ensures a smoothly varying output signal. Unfortunately, we have so far not managed to find a mathematical proof for this boundary condition. Such a proof should follow similar lines as the derivation of the eigenvalue problem in Theorem 1 (see chapter 3 and Appendix A).

It would also be interesting to analyze the properties of solutions that are not translation-invariant, particularly in the light of more complicated receptive field properties such as side- and end-stopping. These effects were also found in the simulations and are inherently not translation-invariant.

A question that cannot be answered within the mathematical framework presented here is what happens if the statistics of the input are not invariant with respect to the transformations at hand. Would the optimal functions for SFA show a different orientation tuning, if the orientation dependence of natural image statistics were taken into account, e.g., by using natural videos as training data? Slowness-based learning of complex cells from natural videos has been done (K. Körding et al., 2004) but to our knowledge not been systematically analyzed from this perspective. Experimentally, it has been shown for cats that an extreme dependence of image statistics on orientation during rearing has a strong impact on the orientation tuning properties of cells in V1 (Hirsch & Spinelli, 1971). More research will be necessary to assess if these influences can be explained in terms of slowness learning. From the theoretical perspective it would be interesting if there is a "unified theory" that captures both the finite-dimensional case presented in chapter 3 and the case considered in this chapter. Such a theory could describe input statistics that are not invariant with respect to the transformations but still capture the restrictions on the function space.

From the perspective of the introductory discussion of the chapter one could argue that any learning rule that aims at explaining the response properties of cells in V1 should, given the maturity of these properties shortly after birth, be able to establish the same receptive field structure from natural images and from retinal waves. Thus, it would be interesting to investigate if complex cell receptive fields can be learned from image sequences that mimic the spatiotemporal structure of retinal waves. Since propagating waves can be interpreted as a "prenatal imitation" of translation in visual scenes, it is likely that slowness learning on these patterns can lead to translation-invariant units with similar response properties as complex cells.

Chapter 6

Slowness and Predictive Coding: An Information-Theoretic Relation

6.1 Introduction

The slowness principle is only one of a range of approaches that have been proposed as candidate mechanisms for learning in sensory systems. An approach that is particularly interesting from the behavioral perspective is predictive coding. The basic idea is that the current state of the environment is probably much less interesting for an animal than the expected state in the near future. In order to catch a fish, a bear needs to anticipate the position of the fish at the moment he strikes. The fish's current position and velocity is useful for estimating its predicted position, so they are behaviorally more relevant than, e.g., its color. From this perspective, aspects of sensory input that are predictive for future events are more important than others. Without this form of prediction, no predator could hunt, no prey could escape its hunter and no tennis player could win a match.

The prediction of future events, in particular of *expected* or *predicted rewards*, plays a central role in reinforcement learning (Sutton & Barto, 1998; Schultz & Dickinson, 2000). In the context of unsupervised learning of sensory processing, predictive coding has been proposed to be the function of a set of inhibitory circuits that mediate surround inhibition in the visual system (Srinivasan et al., 1982; Rao & Ballard, 1999; Hosoya et al., 2005). The idea in these studies is that the visual system predicts input signals based on an internal world model and that the signals exchanged between visual areas are differences between the actual and the predicted input rather than visual features as usually assumed in feed-forward models of visual processing. This approach belongs to the class of models that aim at reducing redundancies in the presentation of sensory data. Sensory data that can be predicted on the basis of correlations with other data carries no new information, so an efficient representation should discard those data and concentrate on those that are not predicted by previous data.

Most of these models concentrate on removing redundancies in the spatial domain, e.g., on the observation that the light intensity at a given retinal position can be predicted from the intensity at neighboring positions (Srinivasan et al., 1982) or that the presence of a bar in a given image patch is a predictor for a collinear bar in a neighboring image

patch (Rao & Ballard, 1999). The associated physiological mechanisms for redundancy reduction are surround inhibition in the retina and end-stopping in hyper-complex cells in V1. The temporal structure of receptive fields in the retina have been interpreted as a redundancy reduction in the temporal domain (Hosoya et al., 2005).

These models for visual processing all possess a structure that conveys the predicted input signal to a given sensor neuron, which then calculates the difference between the actual input and the expected input. But how are these predictions calculated and how can this computation be learned? Here, we will introduce an information-theoretic approach to this question that relies on the information bottleneck technique as introduces by Tishby et al. (1999).

As discussed earlier, slowness – as reflected by the Δ-value of a signal – is related to the autocorrelation time of the signal (see section 4.1.3). Thus, for a slow signal, the current value will in general be more correlated with a future value than for a quickly varying signal. This indicates that slowly varying aspects of the input data are more predictable. Based on this observation, it has been speculated that there may be a relation between the slowness principle and predictive coding (Shaw, 2006).

In this chapter, we show that there is indeed such a relation. Under the assumptions of linear processing and reversible Gaussian input data, we will show that the optimal input-output functions for a temporally local version of predictive coding are the same as those of SFA, apart from information theoretic factors. It will become clear, however, that the relation between slowness and predictive coding has its limitations, because it exists only under assumptions that abolish central components of the original motivation of predictive coding.

The research presented in this chapter was done in cooperation with Felix Creutzig. Most of the results will shortly be published in (Creutzig & Sprekeler, 2008).

6.2 The Gaussian Information Bottleneck

The Information Bottleneck

By construction, the information bottleneck is a supervised learning technique: The goal is to extract aspects of data $\mathbf{x}$ that are informative about a set of *relevance data* $\mathbf{r}$. At the same time, aspects of data $\mathbf{x}$ that are not informative about $\mathbf{r}$ should be discarded. Functionally, the relevance data $\mathbf{r}$ play the role of a supervision signal. This task is achieved by compressing the input data $\mathbf{x}$ into a new representation $\mathbf{y}$ while preserving as much information as possible about $\mathbf{r}$. A formalization of this problem has been given in the form of an optimization problem (Tishby et al., 1999):

$$\text{minimize } \mathcal{L} : \mathcal{L} \equiv I(\mathbf{x};\mathbf{y}) - \beta I(\mathbf{y};\mathbf{r})\,, \tag{6.1}$$

where

$$I(\mathbf{x},\mathbf{y}) = \int p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) \ln \left(\frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{y}}(\mathbf{y})} \right) \mathrm{d}^N x \, \mathrm{d}^M y \tag{6.2}$$

denotes the mutual information between the signals $\mathbf{x}$ and $\mathbf{y}$.

The first term in the objective (6.1) aims at minimizing the complexity of the mapping between $\mathbf{x}$ and $\mathbf{y}$ while the second term increases the accuracy of the representation of $\mathbf{r}$ by $\mathbf{y}$. The relative importance of the two objectives is controlled by the trade-off parameter β.

The information bottleneck method has been applied to a number of different problems, e.g., for document clustering (Slonim & Tishby, 2000), neural code analysis (Dimitrov & Miller, 2001), gene expression analysis (Friedman et al., 2001) and extraction of speech features (Hecht & Tishby, 2005). Here, we will use it to state an information theoretic formulation of predictive coding.

Mutual Information for Gaussian Variables

In general, the estimation of mutual information of two signals is a rather difficult task, because mutual information is defined in terms of joint probability densities. Unfortunately, the estimation of probability densities from a limited set of samples is an ill-posed problem (Vapnik, 2000). Therefore, it is usually more reliable to estimate mutual information directly from the data, e.g., by expansions in terms of the cumulants of the data (Gram-Charlier or Edgeworth expansion (Barndorff-Nielsen & Cox, 1989; Blaschke & Wiskott, 2004)).

The simplest version of such an expansion is obtained by taking only second order statistics into account. The associated cumulants are the mean and the covariance matrix of the data. In essence, this is equivalent to approximating the probability distribution of a set of data $\mathbf{z}$ by a Gaussian

$$p_{\mathbf{z}}(\mathbf{z}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{C}_{\mathbf{z}}|}} \exp\left(-\frac{1}{2}\mathbf{z}^T \mathbf{C}_{\mathbf{z}}^{-1} \mathbf{z}\right) , \tag{6.3}$$

where $\mathbf{C}_{\mathbf{z}} := \langle \mathbf{z}\mathbf{z}^T \rangle_t$ denotes the covariance matrix of $\mathbf{z}$, $|\mathbf{C}_{\mathbf{z}}|$ its determinant and N the dimensionality of the variable $\mathbf{z}$. For simplicity, here and in the remainder of the chapter we will assume that the data have zero mean.

One of the advantages of the Gaussian approximation is that the mutual information between two variables $\mathbf{x}$ and $\mathbf{y}$ with a Gaussian joint distribution $p_{\mathbf{x},\mathbf{y}}$ can simply be expressed in terms of covariance matrices. This can be seen by first writing the mutual information in terms of entropies

$$I(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{x}|\mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}) . \tag{6.4}$$

For Gaussian variables, the entropy h can be calculated analytically and takes a simple form:

$$h(\mathbf{z}) := -\int p_{\mathbf{z}} \ln(p_{\mathbf{z}}) \mathrm{d}^N z = \frac{1}{2} \ln\left((2\pi e)^N |\mathbf{C}_{\mathbf{z}}|\right) , \tag{6.5}$$

The conditional entropy $h(\mathbf{x}|\mathbf{y})$ for Gaussian variables is given by

$$h(\mathbf{y}|\mathbf{x}) := -\int p_{\mathbf{x}} \left[\int p_{\mathbf{y}|\mathbf{x}} \ln(p_{\mathbf{y}|\mathbf{x}}) \, \mathrm{d}^N y\right] \mathrm{d}^N x = \frac{1}{2} \ln\left((2\pi e)^N |\mathbf{C}_{\mathbf{y}|\mathbf{x}}|\right) , \tag{6.6}$$

where the conditional covariance matrix $\mathbf{C}_{\mathbf{y}|\mathbf{x}}$ can be expressed in terms of cross-covariance matrices $\mathbf{C}_{\mathbf{y};\mathbf{x}} := \langle \mathbf{y}\mathbf{x}^T \rangle_t$ by means of Schur's complement formula (Magnus & Neudecker, 1988)

$$\mathbf{C}_{\mathbf{y}|\mathbf{x}} := \left\langle \left\langle (\mathbf{y} - \langle \mathbf{y} \rangle_{\mathbf{y}|\mathbf{x}})(\mathbf{y} - \langle \mathbf{y} \rangle_{\mathbf{y}|\mathbf{x}})^T \right\rangle_{\mathbf{y}|\mathbf{x}} \right\rangle_{\mathbf{x}} \tag{6.7}$$

$$= \mathbf{C}_{\mathbf{y}} - \mathbf{C}_{\mathbf{y};\mathbf{x}} \mathbf{C}_{\mathbf{x}}^{-1} \mathbf{C}_{\mathbf{x};\mathbf{y}} . \tag{6.8}$$

These expressions yield a relatively simple expression for the mutual information

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(\ln(|\mathbf{C}_{\mathbf{x}}| - \ln|\mathbf{C}_{\mathbf{x}|\mathbf{y}}|\right) = \frac{1}{2}\left(\ln(|\mathbf{C}_{\mathbf{y}}| - \ln|\mathbf{C}_{\mathbf{y}|\mathbf{x}}|\right) . \tag{6.9}$$

Mutual Information for a Linear Mapping

In the following, we will consider the simple case where the compressed representation $\mathbf{y}$ is a linear function of the input signal $\mathbf{x}$

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \xi\,, \tag{6.10}$$

with some matrix $\mathbf{A}$ and a Gaussian white noise term ξ. The noise term is introduced for reasons of regularization, so that the mutual information between $\mathbf{x}$ and $\mathbf{y}$ is not divergent. We assume that the noise has unit covariance, i.e., $\langle \xi \xi^T \rangle = \mathbf{I}$, where $\mathbf{I}$ denotes the unit matrix. This can be done without loss of generality, because the mutual information is invariant with respect to arbitrary coordinate changes of either of the variables. If the noise is not normalized, but has covariance matrix $\mathbf{C}_\xi$, we can always perform a linear transformation of the output signal $\mathbf{y}$ such that the objective function $\mathcal{L}$ remains the same and the noise is normalized (Chechik et al., 2005). This linear transformation essentially rescales the signals to recover the signal-to-noise ratio for the case of normalized noise, as discussed in section 6.4.

For such a linear mapping, the mutual information $I(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ can be expressed in terms of the covariance matrix $\mathbf{C}_\mathbf{x}$ of the input data $\mathbf{x}$ and the matrix $\mathbf{A}$:

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left(\ln|\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{C}_\xi| - \ln|\mathbf{C}_\xi|\right) = \frac{1}{2}\ln|\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I}|\,. \tag{6.11}$$

The latter is true for $\mathbf{C}_\xi = \mathbf{I}$.

The other mutual information of interest for the information bottleneck is that between the relevance variable $\mathbf{r}$ and the output signal $\mathbf{y}$. For a linear relationship between $\mathbf{x}$ and $\mathbf{y}$ and for the case that the joint distribution of $\mathbf{r}$ and $\mathbf{x}$ is Gaussian, this information can also be expressed in terms of covariance matrices

$$\begin{aligned} I(\mathbf{y}, \mathbf{r}) &= h(\mathbf{y}) - h(\mathbf{y}|\mathbf{r}) && (6.12)\\ &= \frac{1}{2}\ln|\mathbf{C}_\mathbf{y}| - \frac{1}{2}\ln|\mathbf{C}_{\mathbf{y}|\mathbf{r}}| && (6.13)\\ &= \frac{1}{2}\ln|\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I}| - \frac{1}{2}\ln|\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I}|\,. && (6.14)\end{aligned}$$

The Gaussian Information Bottleneck

Using these expressions, we can rewrite the objective function for the information bottleneck purely in terms of covariance matrices:

$$\mathcal{L} = \frac{1-\beta}{2}\ln|\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I}| + \frac{\beta}{2}\ln|\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I}|\,. \tag{6.15}$$

The advantage of the Gaussian information bottleneck is that the optimal matrix $\mathbf{A}$ can be constructed in terms of the solutions of a generalized eigenvalue problem (Chechik et al., 2005, for a proof see Appendix A.3):

$$\mathbf{A}(\beta) = \begin{cases} [\mathbf{0}; \ldots; \mathbf{0}] & \text{for} \quad 0 \le \beta \le \beta_1^c \\ \left[\alpha_1 \mathbf{w}_1^T; \mathbf{0}; \ldots; \mathbf{0}\right] & \text{for} \quad \beta_1^c \le \beta \le \beta_2^c \\ \left[\alpha_1 \mathbf{w}_1^T; \alpha_2 \mathbf{w}_2^T; \mathbf{0}; \ldots; \mathbf{0}\right] & \text{for} \quad \beta_2^c \le \beta \le \beta_3^c \\ \vdots \end{cases} \tag{6.16}$$

where $\mathbf{0}$ denotes an m-dimensional row vector of zeros and semicolons separate the rows in the matrix $\mathbf{A}(\beta)$. $\mathbf{w}_i$ and λ_i (assume $\lambda_1 \leq \lambda_2 \leq \ldots$) are the eigenvectors and eigenvalues of the generalized eigenvalue problem

$$\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{w}_i = \lambda_i \mathbf{C}_{\mathbf{x}}\mathbf{w}_i \,. \tag{6.17}$$

The eigenvectors $\mathbf{w}_i$ are assumed to be normalized such that the signal $\mathbf{w}_i^T\mathbf{x}$ has unit variance , i.e., $\mathbf{w}_i^T\mathbf{C}_{\mathbf{x}}\mathbf{w}_i = 1$. The coefficients α_i are given by

$$\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i}} \tag{6.18}$$

and $\beta_i^c = \frac{1}{1-\lambda_i}$ are *critical values* for the trade-off parameter β.

The eigenvalues λ_i are guaranteed to be real and non-negative, since both $\mathbf{C}_{\mathbf{x}}$ and $\mathbf{C}_{\mathbf{x}|\mathbf{r}}$ are positive semidefinite. An easy way of seeing this is by multiplying (6.17) from the left with $\mathbf{w}_i^T$:

$$0 \overset{(6.7)}{\leq} \mathbf{w}_i^T\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{w}_i \overset{(6.17)}{=} \lambda_i \underbrace{\mathbf{w}_i^T\mathbf{C}_{\mathbf{x}}\mathbf{w}_i}_{=1} = \lambda_i \,. \tag{6.19}$$

The key observation is that with increasing trade-off parameter β additional eigenvectors enter the matrix $\mathbf{A}$ whenever β reaches a critical value. Eigenvectors with small eigenvalue appear first, i.e., for the smallest β-values. Intuitively, a small eigenvalues means that given the value for $\mathbf{r}$, $\mathbf{x}$ has a small conditional variance when projected on the corresponding eigenvector. Clearly, this is the case for those directions which are most strongly correlated with $\mathbf{r}$, i.e., for directions in $\mathbf{x}$ that are informative about $\mathbf{x}$.

The eigenvectors are orthogonal with respect to the scalar product $(\mathbf{w}, \mathbf{w}') := \mathbf{w}^T\mathbf{C}_{\mathbf{x}}\mathbf{w}'$ that is induced by the covariance matrix $\mathbf{C}_{\mathbf{x}}$. Thus, the output signals $\mathbf{y}$ are uncorrelated. For Gaussian variables, decorrelation is equivalent to statistical independence, so whenever a new eigenvector enters the matrix $\mathbf{A}$, a new output signal becomes available that conveys new information about the relevance signal $\mathbf{r}$.

6.3 Predictive Coding as an Information Bottleneck

Predictive Coding

The idea of predictive coding is that an organism needs to extract information from its sensory input that is predictive for future events. Here, we will present a model for how such a coding scheme could be learned. We assume that there is a set of time-dependent sensory input data $\mathbf{x}_t$. The goal is to store information about sensory signals in the past into some kind of memory or internal state that provides information about expected sensory information in the future. Of course, the trivial way of acquiring as much information about the future as possible is store the past as a whole, because it contains all the information that is available. Due to possible storage and/or efficiency constraints, it is more sensible, however, to store only those aspects of past sensory inputs that are informative about the future and discard uninformative aspects. This problem has the same trade-off structure as the information bottleneck: The task is to compress data about the past while preserving as much information about the future as possible. Formally, the past plays the role of the input signal $\mathbf{x}$ and the future is the relevance

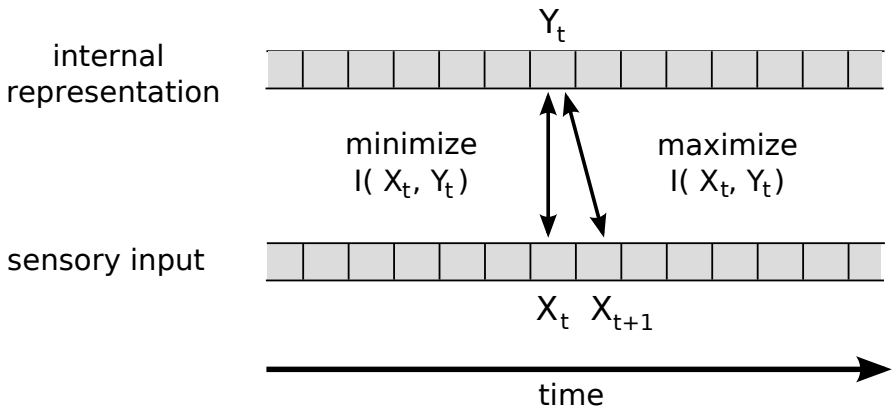


Figure 6.1: **Local predictive coding.** The sensory system compresses information of the current input $\mathbf{x}_t$ into $\mathbf{y}_t$ such that the mutual information between $\mathbf{y}_t$ and the next input $\mathbf{x}_t$ is maximized.

signal $\mathbf{r}$ that defines, which aspects of $\mathbf{x}$ are of interest. We can thus model the problem of predictive coding as an information bottleneck problem:

$$\text{minimize } \mathcal{L} : \mathcal{L} equiv I(\text{past}; \text{internal state}) - \beta\, I(\text{internal state}; \text{future})\,. \tag{6.20}$$

Obviously, the internal state cannot contain more information about the future than about the past, so that for $\beta \leq 1$, the objective function $\mathcal{L}$ is positive: $\mathcal{L} \geq 0$. In this case, $\mathcal{L}$ is optimized by the trivial solution, where the internal state does not contain any information at all, because then $\mathcal{L} = 0$. Thus, to obtain non-trivial solutions, the trade-off parameter should be chosen such that $1 < \beta < \infty$.

Note that in contrast to the original formulation of the information bottleneck, this formulation of predictive coding is an unsupervised learning principle, because the relevance signal is provided by the input data and not by some external signal.

Local Predictive Coding

The prediction problem in its general form (6.20) is of very high complexity, because both the past and the future of the sensory input are infinite-dimensional. Here, we will consider a simplified version of the full prediction problem by restricting the past to the sensory signal $\mathbf{x}_t$ at a single moment t in time and the future to the sensory signal $\mathbf{x}_{t+1}$ one single time step in the future. The restriction of the problem to two time steps would not be an approximation if the sensory data were a first order Markov process, so one could consider this ansatz a Markov approximation of the full prediction problem. Because of the temporal locality, we will refer to this approach as *Local Predictive Coding* (LPC). The associated objective function is given by

$$\mathcal{L} = I(\mathbf{x}_t, \mathbf{y}_t) - \beta I(\mathbf{y}_t, \mathbf{x}_{t+1}) \tag{6.21}$$

and the structure of the problem is illustrated in Figure 6.1.

To further simplify the problem, we assume that the input signal $\mathbf{x}_t$ is Gaussian and that the output signal (or internal state) $\mathbf{y}_t$ is generated by a noisy linear transformation $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \xi_t$. For simplicity, we will assume that the noise ξ is temporally white, i.e., $\langle \xi_t \xi_{t+\tau}^T \rangle_t = 0$ for $\tau \neq 0$, and normalized, i.e., $\langle \xi_t \xi_t^T \rangle_t = \mathbf{I}$.

With these assumptions, the problem has the structure of the Gaussian information bottleneck, so the optimal matrix $\mathbf{A}$ consists of appropriately weighted solutions of the generalized eigenvalue equation

$$\mathbf{C}_{\mathbf{x}_t|\mathbf{x}_{t+1}}\mathbf{w}_i = \lambda_i \mathbf{C}_{\mathbf{x}_t}\mathbf{w}_i\,. \tag{6.22}$$

The first eigenvectors $\mathbf{w}_i$ that appear in the optimal matrix $\mathbf{A}$ are those, for which the variance of $\mathbf{w}_i^T\mathbf{x}_t$ is small given $\mathbf{x}_{t+1}$. One way of achieving this is by choosing directions in the input that vary slowly, because for these directions $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ have similar values, i.e., $\mathbf{x}_t$ has a small conditional variance for fixed $\mathbf{x}_{t+1}$. This indicates a possible connection between slowness learning and predictive coding. In the following, we will put this connection on rigorous mathematical grounds and analyze the restrictions of such a relation.

6.4 Slow Feature Analysis and Local Predictive Coding

Mathematical Analysis

To establish a connection between slow feature analysis and local predictive coding, let us first reconsider the eigenvalue problem that underlies SFA for the linear case. Given the training data $\mathbf{x}_t$, the weight vectors $\mathbf{w}_i$ of the optimal linear functions $g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x}$ are given by the solutions of the eigenvalue problem (2.10)

$$\mathbf{C}_{\dot{\mathbf{x}}_t}\mathbf{w}_i = \Delta_i \mathbf{C}_{\mathbf{x}_t}\mathbf{w}_i\,, \tag{6.23}$$

where $\mathbf{C}_{\dot{\mathbf{x}}_t} := \langle \dot{\mathbf{x}}_t \dot{\mathbf{x}}_t^T \rangle_t \overset{(2.7)}{=} \dot{\mathbf{C}}$ is the covariance matrix of the temporal derivative $\dot{\mathbf{x}}_t$ of $\mathbf{x}_t$ and $\mathbf{C}_{\mathbf{x}_t} = \langle \mathbf{x}_t \mathbf{x}_T^T \rangle_t$ is the covariance matrix of $\mathbf{x}_t$. To avoid confusion with the eigenvalue appearing in LPC, we denote the eigenvalues for SFA with Δ_i.

For temporally discrete data $\mathbf{x}_t$, the temporal derivative is usually approximated by a temporal difference: $\dot{\mathbf{x}}_t \approx \mathbf{x}_{t+1} - \mathbf{x}_t$. With this approximation, it is straightforward to show that the covariance matrix of the temporal derivative can be written in terms of the cross covariance of $\mathbf{x}_t$ for successive time steps (Blaschke et al., 2006):

$$\mathbf{C}_{\dot{\mathbf{x}}_t} = \mathbf{C}_{\mathbf{x}_t} + \mathbf{C}_{\mathbf{x}_{t+1}} - \left[\mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} + \mathbf{C}_{\mathbf{x}_{t+1};\mathbf{x}_t}\right] =: 2\left[\mathbf{C}_{\mathbf{x}_t} - \mathbf{C}^+\right]\,. \tag{6.24}$$

Here, $\mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} = \langle \mathbf{x}_t \mathbf{x}_{t+1}^T \rangle_t$ denotes the cross covariance matrix of $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$. Note that $\mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} = \mathbf{C}_{\mathbf{x}_{t+1};\mathbf{x}_t}^T$, so only the symmetric component $\mathbf{C}^+ := \mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} + \mathbf{C}_{\mathbf{x}_{t+1};\mathbf{x}_t}$ of the cross covariances plays a role in the eigenvalue problem of SFA. Because $\mathbf{C}^+$ is invariant with respect to time reversal, SFA can only capture reversible aspects of the dynamics of $\mathbf{x}_t$. In contrast, for predictive coding, temporal irreversibility should be of utmost importance. A direct link between local predictive coding and SFA can thus only be provided for reversible signals. We will further address this issue in the discussion and assume in the remainder of the chapter that the input signals $\mathbf{x}_t$ have temporally reversible statistics. This implies in particular that $\mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} = \mathbf{C}_{\mathbf{x}_{t+1};\mathbf{x}_t} = \mathbf{C}^+$.

The generalized eigenvalue equation (6.22) for local predictive coding is structurally very similar to the eigenvalue equation for SFA. To show that they have the same eigenvectors, we express the conditional covariance matrix $\mathbf{C}_{\mathbf{x}_t|\mathbf{x}_{t+1}}$ in terms of the discretized

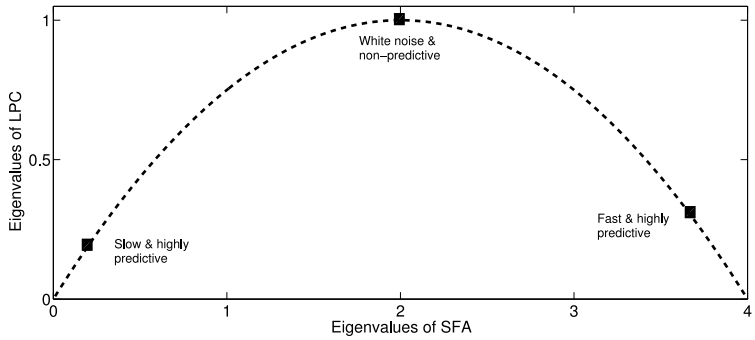


Figure 6.2: **Relationship between eigenvalues of slow feature analysis and local predictive coding.** For discrete time series, the dependence of the LPC eigenvalue on the Δ-value of the signal is not monotonic. Thus, fast components can be equally predictive as slow components, if successive data point are anti-correlated. Only white noise, with a Δ-value of $\Delta = 2$, is completely non-predictive.

covariance matrix $\mathbf{C}_{\dot{\mathbf{x}}_t}$ of the temporal derivative:

$$\begin{aligned}
\mathbf{C}_{\mathbf{x}_t|\mathbf{x}_{t+1}} &\overset{(6.8)}{=} \mathbf{C}_{\mathbf{x}_t} - \mathbf{C}_{\mathbf{x}_t;\mathbf{x}_{t+1}} \mathbf{C}_{\mathbf{x}_{t+1}}^{-1} \mathbf{C}_{\mathbf{x}_{t+1};\mathbf{x}_t} \\
&= \mathbf{C}_{\mathbf{x}_t} - \mathbf{C}^{+} \mathbf{C}_{\mathbf{x}_t}^{-1} \mathbf{C}^{+} \\
&\quad \text{(for reversible signals } \mathbf{x}_t\text{)} \\
&\overset{(6.24)}{=} \mathbf{C}_{\mathbf{x}_t} - (\mathbf{C}_{\mathbf{x}_t} - \tfrac{1}{2}\mathbf{C}_{\dot{\mathbf{x}}_t}) \mathbf{C}_{\mathbf{x}_t}^{-1} (\mathbf{C}_{\mathbf{x}_t} - \tfrac{1}{2}\mathbf{C}_{\dot{\mathbf{x}}_t}) \\
&= \mathbf{C}_{\dot{\mathbf{x}}_t} - \tfrac{1}{4}\mathbf{C}_{\dot{\mathbf{x}}_t} \mathbf{C}_{\mathbf{x}_t}^{-1} \mathbf{C}_{\dot{\mathbf{x}}_t} \,.
\end{aligned}$$

Now let $\mathbf{w}_i$ be an eigenvector for the SFA eigenvalue problem (6.23) with eigenvalue Δ_i. To show that $\mathbf{w}_i$ is also a solution to the LPC eigenvalue problem, we apply $\mathbf{C}_{\mathbf{x}_t|\mathbf{x}_{t+1}}$ to the SFA eigenvector $\mathbf{w}_i$

$$\begin{aligned}
\mathbf{C}_{\mathbf{x}_t|\mathbf{x}_{t+1}} \mathbf{w}_i &\overset{(6.25)}{=} \left[\mathbf{C}_{\dot{\mathbf{x}}_t} - \tfrac{1}{4}\mathbf{C}_{\dot{\mathbf{x}}_t} \mathbf{C}_{\mathbf{x}_t}^{-1} \mathbf{C}_{\dot{\mathbf{x}}_t}\right] \mathbf{w}_i \\
&\overset{(6.23)}{=} \left[\Delta_i - \tfrac{1}{4}\Delta_i^2\right] \mathbf{C}_{\mathbf{x}_t} \mathbf{w}_i \\
&= \lambda_i \mathbf{C}_{\mathbf{x}_t} \mathbf{w}_i \,.
\end{aligned} \tag{6.25}$$

Clearly, the eigenvectors of SFA are also eigenvectors to the LPC problem, with eigenvalues

$$\lambda_i = \frac{1}{4}\Delta_i \left(4 - \Delta_i\right) \,. \tag{6.26}$$

Slowness vs. Predictability for Discrete Signals

The relation between the eigenvalues of SFA and LPC is illustrated in Figure 6.2. A curious observation is that the "choice" of eigenvectors may differ for LPC and SFA, because the relation between the eigenvalues is not monotonic. Fast components with $\Delta > 2$ may have smaller LPC eigenvalues than slower ones. The underlying reason is the following: $\Delta = 2$ corresponds to white noise, i.e., values for successive time steps are uncorrelated. For $\Delta > 2$, values for successive time steps become *anti-correlated*. Although this implies a quickly varying signal, successive values are nevertheless correlated, so the current value of the signal is predictive for the next. A nice illustration is the case of $\Delta = 4$, which can only be implemented by an alternating output signal $y_t = (-1)^t$. In

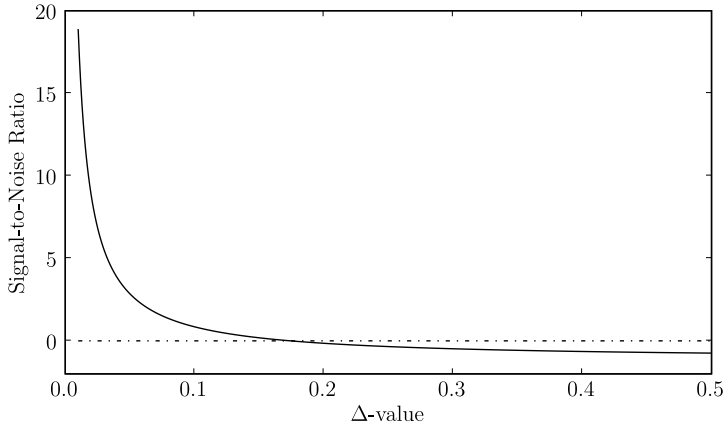


Figure 6.3: **Signal-to-Noise Ratio of the LPC components as a function of the Δ-value.** The dependence of the signal-to-noise ratio for $\beta = 1.2$. Note the zero crossing at $\Delta = 2{-}2/\sqrt{\beta} \approx 0.17$. Signals with Δ-values above this crossing point do not enter the solution of LPC, the associated eigenvectors are discarded.

this case, the signal of course varies quickly, but knowing the value at time t uniquely determines the value at time $t+1$, so that this signal is perfectly predictable.

It should be noted that this difference between SFA and LPC is merely due to the temporal discretization of the input. The presence of signal components that vary "faster" than white noise could be an indication that the sampling rate is too low. For sufficiently well-sampled signals, we expect that signals with $\Delta > 2$ do not occur[1], so that the relation between the eigenvalues is monotonic and the eigenvectors for LPC and SFA will be the same.

Information-Theoretic Weights: Variance vs. Signal-to-Noise Ratio

Another difference between SFA and LPC is that LPC introduces a weight α_i for the extracted components. The optimal matrix $\mathbf{A}(\beta)$ contains the same eigenvectors as SFA, but the variance of the associated output signals $y_i = \alpha_i \mathbf{w}_i^T \mathbf{x}_t + \xi_{i,t}$ depends on the trade-off parameter β and the eigenvalue λ_i or respectively Δ_i:

$$\mathrm{var}(y_i) = \alpha_i^2 \overset{(6.18)}{=} \frac{\beta(1-\lambda_i)-1}{\lambda_i} \overset{(6.26)}{=} \frac{\beta(\Delta_i-2)^2-4}{4-(\Delta_i-2)^2} \overset{\Delta_i \ll 1}{\approx} \frac{\beta-1}{\Delta_i} - \beta\,. \tag{6.27}$$

Note that the variance of the output signals depends on the deviation of their Δ-value from 2, i.e., how strongly the signals differs from white noise. For well-sampled signals, the Δ-value will typically be much smaller than one, which allows to approximate the variance by a Taylor expansion of first order. Qualitatively, the variance of the signal is then inversely proportional to the Δ-value.

The assignment of different variances to different signals implies that different signals have different signal-to-noise ratios (cf. equation (6.10)). In fact, it is more precise to

[1]Note that this does not imply that the signals may not have frequency components that vary with $\Delta > 2$. Such frequency components will surely be present, in particular in the presence of noise. Rather, we assume that there is no linear combination of the input data with $\Delta > 2$.

interpret the weighting factors α_i^2 as signal-to-noise ratios rather than as variances, because of the assumption that the noise ξ has unit covariance matrix. For arbitrary noise with a covariance matrix $\mathbf{C}_\xi$, it has been shown that an optimal matrix $\tilde{\mathbf{A}}$ is given by $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{V}\mathbf{A}$ (Chechik et al., 2005). Here, $\mathbf{A}$ is the solution for normalized noise and $\mathbf{D}$ and $\mathbf{V}$ are the diagonal and the orthogonal matrix containing the eigenvalues and eigenvectors of $\mathbf{C}_\xi = \mathbf{V}\mathbf{D}\mathbf{V}^T$, respectively. Geometrically, the solution for non-normalized noise can thus be constructed by first rotating the solution for normalized noise such that its components lie on the principal axes $\mathbf{V}$ of the noise and then rescaling them by the inverse noise amplitude in this direction. The resulting signals have the same signal-to-noise ratio as in the case of normalized noise. Thus, the factors α_i^2 are the signal-to-noise ratios of the optimal signal components, independent of the correlation structure of the noise.

In contrast to SFA, the number of output signals for LPC is not explicitly chosen by hand. Instead, the number of output signals with non-vanishing variance is implicitly determined by the choice of the trade-off parameter β. This is reflected by the fact that for given β, signals with a Δ-value above a threshold $\Delta_{\max} = 2 \pm 2/\beta$ would be accredited a negative signal-to-noise ratio (see equation (6.27) and Figure 6.3), which is of course not possible. Instead, these signals are neglected, they do not enter the solution of LPC at all (cf. equation (6.16) for the optimal matrix of the Gaussian information bottleneck).

Finally, the objective function for LPC is invariant with respect to arbitrary orthogonal transformations of the output signals. Thus, in contrast to SFA, LPC introduces neither an order of the solutions nor a clear separation in that signals with different Δ-values correspond to different output signals. The output signals are arbitrary mixtures of the rescaled output signals of SFA.

6.5 Discussion

In this chapter, we have provided an information-theoretic relation between slow feature analysis and the principle of predictive coding. Such a relation has previously been suggested in other work, e.g., by Shaw (2006), who argued that temporal invariance learning is equivalent to predictive coding if the input signals are generated from Ornstein-Uhlenbeck processes. However, to our knowledge a clear mathematical link had not yet been established.

The link between slowness and predictive coding requires several assumptions, some of which are rather drastic. The assumed Gaussianity of the input signals is probably the least problematic assumption. By construction, SFA relies on second order statistics[2], so it is not surprising that a one-to-one connection to an information-theoretic technique can only be provided for Gaussian signals, whose probability distributions are uniquely characterized by second-order statistics. From this perspective, it may be possible to interpret SFA as a Gaussian approximation of a more complicated information-theoretic problem.

The restriction of the theory to linear input-output relations is closely related to the restriction of Gaussianity. A nonlinear generalization of LPC can readily be formulated by performing linear LPC after a nonlinear expansion. Unfortunately, however, the statistics of nonlinearly expanded input data are usually highly non-Gaussian, so that a

[2]The nonlinear expansion usually used for SFA of course grants access to higher order statistics of the input data, but for linear SFA, these statistics are invisible.

Gaussian description of the expanded input is a poor approximation. A generalization of LPC to non-Gaussian signals would thus immediately provide the means for nonlinear LPC.

A rather drastic assumption required for the connection between SFA and LPC is that the input signals have temporally reversible statistics. In combination with the assumption that the prediction is done on the basis of the input signal $\mathbf{x}_t$ at a single moment in time, the best guess for the next input can only be the status quo $\mathbf{x}_t$. Thus, predictable signals have to show small deviation from one time step to the next, i.e., they should vary slowly. "Real" predictions, in contrast, should incorporate an extrapolation into the future, either on the basis of several data points for different moments in time (to predict the trajectory of a ball, we need to know its velocity, which cannot be estimated from a single "snapshot") or on the basis of irreversible regularities in the input signals (clouds appear *before* the rain, not afterwards). The combined assumption of temporal locality *and* reversibility allows neither.

For irreversible input statistics, the optimal solutions for SFA and LPC will in general be different, because in contrast to SFA, LPC can exploit irreversibilities. A generalization of LPC to more time steps, which would allow prediction even for reversible statistics, should be straightforward, but is beyond the scope of this work. A dynamical systems approach to the general case of predictive coding that takes the full future and past into account has been presented by Creutzig (2008).

We conclude that although it is satisfying to see that the relation between slowness and prediction that has been made on intuitive grounds can be made explicit, the analysis has shown that the restrictions under which this relation could be established are severe from the perspective of predictive coding. It may be possible to establish a link under more general conditions, but in the light of the discussion above, we believe that this is rather unlikely. Still, it remains interesting to examine to what extent the discussed assumptions are fulfilled for previous applications of predictive coding in the sensory domain and if the learning rules that lead to the predictive elements in these models can be related to SFA.

Part II

On the Biological Plausibility of Slowness Learning

Chapter 7

Slowness: An Objective for Spike-Timing–Dependent Plasticity?

7.1 Introduction

In chapter 4 and 5 we have shown that SFA can serve as a basis for models of cell properties in primary visual cortex (see also Berkes & Wiskott, 2005) and in the hippocampal formation (Franzius, Sprekeler & Wiskott, 2007). This suggests that – on an abstract level – SFA seems to capture aspects of cortical information processing. From the perspective of biological plausibility, however, the SFA algorithm is problematic for several reasons. Firstly, SFA is formulated as a batch learning algorithm, i.e., it gathers information about the whole set of training data (reflected by the covariance matrices of the input signals and their time derivative) and only solves the resulting eigenvalue problem when all input data have been presented. This requires a switching mechanisms that differentiates between periods of information gathering and periods of plasticity. Although such switches may exist in the brain[1], it is simpler to assume that adaptive elements in the cortex undergo continuous changes as the input data arrive. This would speak for online learning rules rather than batch learning. Secondly, the SFA algorithm finds the optimal solutions by solving an eigenvalue problem, an operation that is difficult to imagine for a cortical neuron. Thirdly and last, SFA includes an asymmetric decorrelation constraint: The slowest output signal is not influenced by the decorrelation constraint, while the others have to be uncorrelated to the output signals of slower functions. Why should certain neurons in cortex be treated differently from others?

In this second part of the thesis, we will approach the problem of biological plausibility of slowness learning from two perspectives. In this chapter, we will assess if it is possible to implement the slowness objective in terms of biologically plausible mechanisms, in particular in terms of synaptic plasticity in spiking neurons. In chapter 8, we will give an outlook on how the gradient descent learning rule derived in this chapter can be interpreted in the context of receptive field dynamics and discuss how constraints could be implemented in a biologically more plausible fashion.

[1] For example, reward signals issued by several subcortical brain regions (Schultz & Dickinson, 2000) may play the role of a trigger for intracortical synaptic plasticity (Froemke et al., 2007).

In this chapter we will first consider a continuous model neuron and demonstrate that a modified Hebbian learning rule enables the neuron to learn the slowest (in the sense of SFA) linear combination of its inputs. Apart from providing the basis for the analysis of the spiking model, this section reveals a mathematical link between SFA and the trace learning rule (Földiàk, 1991), another implementation of the slowness principle. We then examine if these findings also hold for a spiking model neuron and find that for a linear Poisson neuron spike-timing–dependent plasticity (STDP) can be interpreted as a gradient-based implementation of the slowness principle.

The research in this chapter was done in collaboration with Christian Michaelis. Large parts of this chapter have been published in (Sprekeler et al., 2007) under an open-access licence.

7.2 Continuous model neuron

7.2.1 Linear Model Neuron and Basic Assumptions

First consider a linear continuous model neuron with an input-output function given by

$$a^{\text{out}}(t) = \sum_{i=1}^{n} w_i \, a_i^{\text{in}}(t) \,, \tag{7.1}$$

with a_i^{in} indicating the input signals, w_i the weights, and a^{out} the output signal. For mathematical convenience, let a_i^{in} and a^{out} be defined on the interval $[-\infty, \infty]$ but differ from zero only on $[0, T]$, which could be the lifetime of the system. We assume that the input is approximately whitened on any sufficiently large interval $[t_a, t_b] \subseteq [0, T]$, i.e. each input signal has approximately zero mean and unit variance and is uncorrelated with other input signals:

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t) \, \mathrm{d}t \approx 0 \,, \qquad \text{(zero mean)} \tag{7.2}$$

$$\frac{1}{T_{ab}} \int_{t_a}^{t_b} a_i^{\text{in}}(t)^2 \, \mathrm{d}t \approx 1 \,, \qquad \text{(unit variance)} \tag{7.3}$$

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t) \, a_{j \neq i}^{\text{in}}(t) \, \mathrm{d}t \approx 0 \,, \qquad \text{(decorrelation)} \tag{7.4}$$

with $T_{ab} = t_b - t_a$.

This can be achieved by a normalization and decorrelation step of the units projecting to the considered unit. Furthermore, we assume that the output is normalized to unit variance, which for whitened input means that the weight vector is normalized to length 1. In an online learning rule, this could be implemented by either an activity-dependent or

a weight-dependent normalization term. Thus, for the output signal we have:

$$\int_{t_a}^{t_b} a^{\text{out}}(t)\,\mathrm{d}t \overset{(7.1,7.2)}{\approx} 0\,, \qquad \text{(zero mean)} \tag{7.5}$$

$$\frac{1}{T_{ab}}\int_{t_a}^{t_b} a^{\text{out}}(t)^2\,\mathrm{d}t \overset{(7.1,7.3)}{\approx} \sum_{i=1}^{n} w_i^2 \overset{!}{=} 1\,. \qquad \text{(unit variance)} \tag{7.6}$$

In the following we will often consider filtered signals. Therefore we introduce abbreviations for the convolution $f \circ g$ and the cross-correlation $f \star g$ of two functions $f(t)$ and $g(t)$:

$$\text{Convolution:} \quad [f \circ g](t) := \int_{-\infty}^{\infty} f(\tau) g(t-\tau)\,\mathrm{d}\tau\,, \tag{7.7}$$

$$\text{Cross-correlation:} \quad [f \star g](t) := \int_{-\infty}^{\infty} f(\tau) g(t+\tau)\,\mathrm{d}\tau\,. \tag{7.8}$$

For convenience, we will often use windowed signals, indicated by a hat

$$\hat{s}(t) := \begin{cases} s(t) & \text{if } t \in [t_a, t_b] \\ 0 & \text{otherwise} \end{cases}\,, \tag{7.9}$$

which allows us to replace the integration of a signal $s(t)$ over $[t_a, t_b]$ by an integration of $\hat{s}(t)$ over $[-\infty, \infty]$. We assume that the interval $[t_a, t_b]$ is long compared to the width of the filters. In this case effects from the integration boundaries are negligible, and we have

$$\int_{t_a}^{t_b} [f \circ s](t) h(t)\,\mathrm{d}t \approx \int_{-\infty}^{\infty} [f \circ \hat{s}](t) h(t)\,\mathrm{d}t\,. \tag{7.10}$$

Similar considerations hold for the cross-correlation (7.8).

Since convolution and cross-correlation are conveniently treated in Fourier space, we repeat the definition of the Fourier transform $\mathcal{F}_s(\nu)$ and the power spectrum $\mathcal{P}_s(\nu)$ of a signal $s(t)$.

$$\text{Fourier transform:} \quad s(t) =: \int_{-\infty}^{\infty} \mathcal{F}_s(\nu)\, e^{2\pi \mathrm{i} \nu t}\,\mathrm{d}\nu \tag{7.11}$$

$$\text{Power spectrum:} \quad \mathcal{P}_s(\nu) := \mathcal{F}_s(\nu)\overline{\mathcal{F}_s}(\nu)\,. \tag{7.12}$$

Throughout the paper, we make the assumption that input signals (and hence also the output signals) do not have significant power above some reasonable frequency ν_{max}.

7.2.2 Reformulation of the Slowness Objective

SFA is based on the minimization of the second moment of the time derivative, $\int \dot{a}^{\text{out}}(t)^2\,\mathrm{d}t$. Even though there are neurons with transient responses to changes in the input, we believe it would be more plausible if we could derive an SFA-learning rule that does not

depend on the time derivative, because it might be difficult to extract, especially for spiking neurons. It is indeed possible to replace the time derivative by a low-pass filtering as follows.

$$\text{minimize} \quad \int_{-\infty}^{\infty} \dot{a}^{\text{out}}(t)^2 \,\mathrm{d}t \tag{7.13}$$

$$= \int_{-\infty}^{\infty} \mathcal{P}_{\dot{a}^{\text{out}}}(\nu) \,\mathrm{d}\nu \quad \text{(because of Parseval's theorem)} \tag{7.14}$$

$$= 4\pi^2 \int_{-\infty}^{\infty} \nu^2 \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu \quad (\text{since } \mathcal{F}_{\dot{s}}(\nu) = 2\pi \mathrm{i} \nu \mathcal{F}_s(\nu)) \tag{7.15}$$

$$\iff \text{maximize} \quad \int_{-\infty}^{\infty} -\nu^2 \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu \tag{7.16}$$

$$\iff \text{maximize} \quad \int_{-\infty}^{\infty} (\nu_{max}^2 - \nu^2) \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu \tag{7.17}$$

$$\left(\text{since } \int_{-\infty}^{\infty} \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu = \int_{-\infty}^{\infty} a^{\text{out}}(t)^2 \,\mathrm{d}t \overset{(7.6)}{\approx} \text{const}\right)$$

$$= \int_{-\infty}^{\infty} \max(0, (\nu_{max}^2 - \nu^2)) \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu \tag{7.18}$$

$$\left(\text{since } \mathcal{P}_{a^{\text{out}}}(\nu) = 0 \text{ for } |\nu| > \nu_{max} \text{ by assumption}\right)$$

$$= \int_{-\infty}^{\infty} \mathcal{P}_{f_{SFA}}(\nu) \mathcal{P}_{a^{\text{out}}}(\nu) \,\mathrm{d}\nu \tag{7.19}$$

$$\left(\text{with } f_{SFA}(t) \text{ such that } \mathcal{P}_{f_{SFA}} = \max(0, (\nu_{max}^2 - \nu^2))\right) \tag{7.20}$$

$$= \int_{-\infty}^{\infty} \left[f_{SFA} \circ a^{\text{out}}\right](t)^2 \,\mathrm{d}t \,. \tag{7.21}$$

Thus, SFA can be achieved either by minimizing the variance of the time derivative of the output signal or by maximizing the variance of the appropriately filtered output signal. Figure 7.1 provides an intuition for this alternative. The filter f_{SFA} is obviously a low-pass filter, as one would expect, with a $(\nu_{max}^2 - \nu^2)$-power spectrum below the limiting frequency ν_{max}. Because the phases are not determined, further assumptions are required to fully determine an SFA filter. However, we will proceed without defining a concrete filter, since it is not required for the considerations below.

7.2.3 Hebbian Learning on Filtered Signals

It is known that standard Hebbian learning under the constraint of a unit weight vector applied to a linear unit maximizes the variance of the output signal. We have seen in the previous section that SFA can be reformulated as a maximization problem for the variance of the low-pass filtered output signal. To achieve this we simply apply Hebbian learning to the filtered input and output signals instead of to the original signals.

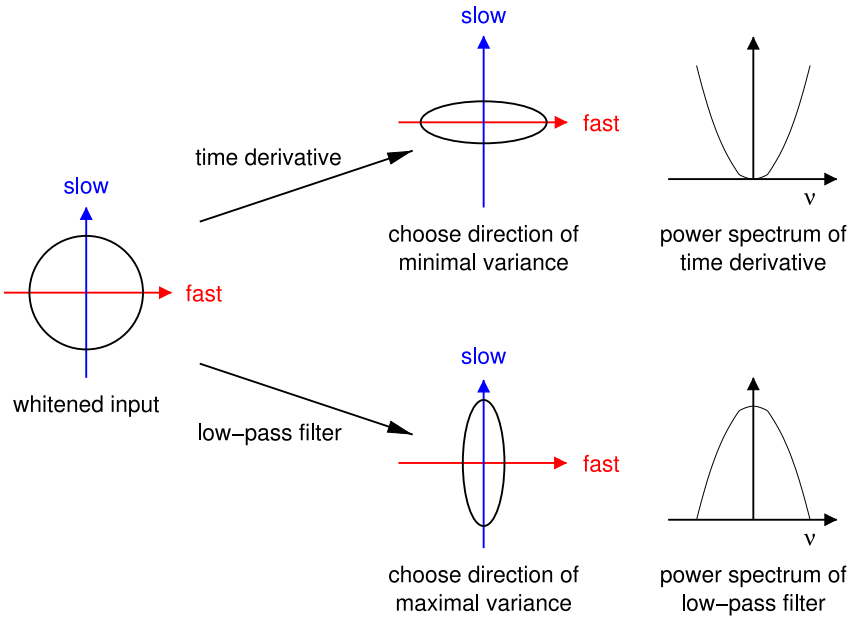


Figure 7.1: **Choosing slow directions of the input**. Finding the direction of least variance in the time derivative of the input (which is part of the SFA algorithm) can be replaced by finding the direction of maximum variance in an appropriately low-pass filtered version of the input signal.

Consider a hypothetical unit that receives low-pass filtered inputs and therefore, because of the linearity of the unit and the filtering, generates a low-pass filtered output

$$[f_{SFA} \circ a^{\text{out}}](t) \overset{(7.1)}{=} \left[f_{SFA} \circ \sum_{i=1}^{n} w_i \, a_i^{\text{in}} \right](t) = \sum_{i=1}^{n} w_i \, [f_{SFA} \circ a_i^{\text{in}}](t) \,, \tag{7.22}$$

where f_{SFA} is the kernel of the linear filter applied. It is obvious that a *filtered Hebbian learning rule*

$$\dot{w}_i = \eta \, [f^{\text{in}} \circ a_i^{\text{in}}](t) \, [f^{\text{out}} \circ a^{\text{out}}](t) \tag{7.23}$$

with $f^{\text{in}} := f^{\text{out}} := f_{SFA}$ maximizes the objective (7.21).

Remember that the input is white (i.e., the a_i^{in} are uncorrelated and have unit variance) and the weight vector is normalized to norm one by some additional normalization rule, so that we know that the output signal a^{out} has the same variance no matter what the direction of the weight vector is. Thus, the filtered Hebbian plasticity rule (together with the normalization rule not specified here) optimizes the slowness objective (7.13) under the constraint (7.6) of unit variance. Figure 7.2 illustrates this learning scheme. It also underlines the necessity for a clear distinction between processing and learning. Although the slowness principle does not allow low-pass filtering as a means of generating slow signals during processing, the learning rule may well make use of low-pass filtered signals in order to detect slowly varying features in the input signal. This distinction will become particularly important for the Poisson model neuron below, as it incorporates an excitatory postsynaptic potential (EPSP) that acts as a low-pass filter during processing. An implementation of the slowness principle in such a system must avoid the system exploiting the EPSP as a means of generating slow signals.

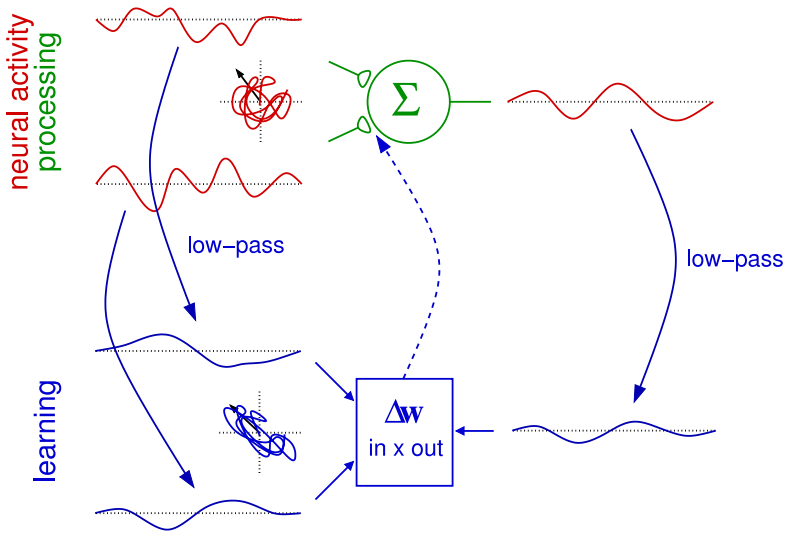


Figure 7.2: **'Filtered Hebbian' learning rule**. Input and output signals are filtered (downward arrows). The weight change is the result of applying the Hebbian learning rule on the filtered signals (square box and upward arrow). Thereby, the variance of the filtered version of the output is maximized without actually filtering the output during processing.

7.2.4 Alternative Filtering Procedures

If learning is slow, the total weight change over a time interval $[t_a, t_b]$ in a synapse can be written as

$$\Delta w_i := \int_{t_a}^{t_b} \dot{w}_i \, \mathrm{d}t \tag{7.24}$$

$$\overset{(7.23)}{=} \eta \int_{t_a}^{t_b} [f^{\mathrm{in}} \circ a_i^{\mathrm{in}}](t) \, [f^{\mathrm{out}} \circ a^{\mathrm{out}}](t) \, \mathrm{d}t \tag{7.25}$$

$$\overset{(7.10)}{\approx} \eta \int_{-\infty}^{\infty} [f^{\mathrm{in}} \circ \hat{a}_i^{\mathrm{in}}](t) \, [f^{\mathrm{out}} \circ \hat{a}^{\mathrm{out}}](t) \, \mathrm{d}t \tag{7.26}$$

$$= \eta \int_{-\infty}^{\infty} [[f^{\mathrm{out}} \star f^{\mathrm{in}}] \circ \hat{a}_i^{\mathrm{in}}](t) \, \hat{a}^{\mathrm{out}}(t) \, \mathrm{d}t \tag{7.27}$$

$$= \eta \int_{-\infty}^{\infty} \hat{a}_i^{\mathrm{in}}(t) \, [[f^{\mathrm{in}} \star f^{\mathrm{out}}] \circ \hat{a}^{\mathrm{out}}](t) \, \mathrm{d}t \tag{7.28}$$

$$= \eta \int_{-\infty}^{\infty} [f^{\mathrm{in}} \star f^{\mathrm{out}}](t) \, [\hat{a}^{\mathrm{out}} \star \hat{a}_i^{\mathrm{in}}](t) \, \mathrm{d}t \, . \tag{7.29}$$

Thus one can either convolve input and output signal with filters f^{in} and f^{out}, respectively, the input signal with $f^{\mathrm{out}} \star f^{\mathrm{in}}$, or the output signal with $f^{\mathrm{in}} \star f^{\mathrm{out}}$. Note that $[f^{\mathrm{in}} \star f^{\mathrm{out}}](t) = [f^{\mathrm{out}} \star f^{\mathrm{in}}](-t)$. One can actually use any pair of filters f^{in} and f^{out} as

long as $f^{\text{in}} \star f^{\text{out}}$ fulfills the condition

$$\mathcal{F}_{f^{\text{in}} \star f^{\text{out}}}(\nu) = \mathcal{P}_{f_{SFA}}(\nu)\,. \tag{7.30}$$

7.2.5 Relation to Other Learning Rules

Hebbian learning on low-pass filtered signals is the basis of several other models for unsupervised learning of invariances (Földiàk, 1991; O'Reilly & Johnson, 1994; Wallis & Rolls, 1997). These models essentially subject the output signal to an exponential temporal filter $f(t) := \theta(t)\gamma \exp(-\gamma t)$ and then use Hebbian learning to associate it with the input signal. Here $\theta(t)$ denotes the Heaviside step function, which is 0 for $t < 0$ and 1 for $t \geq 0$. This learning rule has been named the *trace rule*. The considerations in the last section provide a link between this approach and ours. We simply have to replace $f^{\text{in}}(t)$ with a δ-function and $f^{\text{out}}(t)$ with $f(t)$. Equation (7.29) then takes the form

$$\Delta w_i = \eta \sum_j \left[\int_{-\infty}^{\infty} f(t)\, [\hat{a}_j^{\text{in}} \star \hat{a}_i^{\text{in}}](t)\, \mathrm{d}t \right] w_j\,, \tag{7.31}$$

since the output signal $a^{\text{out}} = \sum_j w_j a_j^{\text{in}}$ is a linear function of the input (see equation (7.1)). In the previously mentioned applications of the trace rule, the statistics of the input signals were always reversible, so we will assume that all correlation functions $[a_i^{\text{in}} \star a_j^{\text{in}}](t)$ are symmetric in time. This implies that only the symmetric component of $f(t)$ is relevant for learning:

$$f^{\text{sym}}(t) := \frac{1}{2}(f(t) + f(-t)) = \frac{\gamma}{2} \exp(-\gamma|t|). \tag{7.32}$$

It is easy to show that the learning rule (7.31) can be interpreted as a gradient ascent on the following objective function:

$$\Psi = \int_{-\infty}^{\infty} f^{\text{sym}}(t)[\hat{a}^{\text{out}} \star \hat{a}^{\text{out}}](t)\, \mathrm{d}t \tag{7.33}$$

$$= \int_{-\infty}^{\infty} \mathcal{F}_{f^{\text{sym}}}(\nu)\mathcal{P}_{\hat{a}^{\text{out}}}(\nu)\, \mathrm{d}\nu\,. \tag{7.34}$$

By comparison with equation (7.20), it becomes clear that the trace rule implements a very similar objective as our model. The only difference is that the power spectrum in equation (7.20) is replaced by the Fourier transform of the filter f^{sym}. Note that in order to be able to interpret Ψ as an objective function, it should be real-valued. The replacement of f with f^{sym} ensures that $\mathcal{F}_{f^{\text{sym}}}$ is real-valued and symmetric, so Ψ is real-valued as well. The Fourier transform of f^{sym} is given by

$$\mathcal{F}_{f^{\text{sym}}}(\nu) = \frac{\gamma}{\gamma^2 + (2\pi\nu)^2}. \tag{7.35}$$

This shows that the only difference between the trace rule and our model lies in the choice of the power spectrum for the low-pass filter. While we are using a parabolic power spectrum with a cutoff (7.20) the trace rule uses a power spectrum with the shape of a Cauchy function (7.35).

From this perspective, one can interpret SFA as a quadratic approximation of the trace rule. To what extent this approximation is valid depends on the power spectra of the input signals. If most of the input power is concentrated at low frequencies, where the power spectrum resembles a parabola, the learning rules can be expected to learn very similar weight vectors. In fact, any Hebbian learning rule that leads to an objective function of the shape of equation (7.20) with a low-pass filtering spectrum in the place of $\mathcal{P}_{f_{SFA}}$ essentially implements the slowness principle, as among signals with the same variance, it will favor slower ones.

7.3 Spiking model neuron

Real neurons do not transmit information via a continuous stream of analog values like the model neuron considered in the previous section, but rather emit action potentials that carry information by means of their rate and probably also by their exact timing, a fact we will not consider here. How can the model developed so far be mapped onto this scenario?

7.3.1 The Linear Poisson Neuron

Again, we restrict our analysis to a simple case by modeling the spike train signals by inhomogeneous Poisson processes. Note that at this point, we restrict our analysis to a rate code, thus neglecting possible coding paradigms that rely on precise timing of spikes.

To generate the input spike trains, we first add sufficiently large constants c_i^{in} to the continuous and zero-mean signals $a_i^{\text{in}}(t)$ to turn them into strictly positive signals that can be interpreted as rates

$$r_i^{\text{in}}(t) := c_i^{\text{in}} + a_i^{\text{in}}(t)\,. \tag{7.36}$$

The constants c_i^{in} represent mean firing rates, which are modulated by the input signals a_i^{in}. From the input rates $r_i^{\text{in}}(t)$ we then derive inhomogeneous Poisson spike trains $S_i^{\text{in}}(t)$ drawn from ensembles E_i^{in} such that

$$\langle S_i^{\text{in}}(t)\rangle_{E_i^{\text{in}}} = r_i^{\text{in}}(t)\,, \tag{7.37}$$

where $\langle\cdot\rangle_{E_i^{\text{in}}}$ denotes the average over the ensemble E_i^{in}.

The output rate is modeled as a weighted sum over the input spike trains convolved with an EPSP $\epsilon(t)$ plus a baseline firing rate r_0, which ensures that the output firing rate remains positive. This is necessary as we allow inhibitory synapses, (i.e., negative weights).

$$m(t) := r_0 + \sum_{i=1}^{n} w_i\,[\epsilon \circ S_i^{\text{in}}](t)\,. \tag{7.38}$$

Note that in this scheme, the EPSP reflects the change in the postsynaptic firing probability rather than a change in the membrane potential. Ideally, it includes all delay effects in neuronal transmission.

The output of this spiking neuron is yet another inhomogeneous Poisson spike train $S^{\text{out}}(t)$ drawn from an ensemble E^{out} given a realization of the input spike-trains S_i^{in} such that

$$\langle S^{\text{out}}(t)\rangle_{E^{\text{out}}|\{S_i^{\text{in}}\}} = m(t)\,. \tag{7.39}$$

It should be noted that not only is the output spike train $S^{\text{out}}(t)$ stochastic in this model, but also the underlying output rate $m(t)$, which is a function of the stochastic variables $S_i^{\text{in}}(t)$ and generally differs for each realization of the input. This is the reason why the input and output spike trains are not statistically independent. However, due to the linearity of the model neuron, the output rate is still simply

$$r^{\text{out}}(t) := \langle S^{\text{out}}(t) \rangle_{E^{\text{in}},E^{\text{out}}} \tag{7.40}$$

$$\overset{(7.39,7.38,7.37)}{=} r_0 + \sum_{i=1}^{n} w_i \left[\epsilon \circ r_i^{\text{in}}\right](t) \tag{7.41}$$

$$\overset{(7.36)}{=} \underbrace{r_0 + \sum_{i=1}^{n} w_i c_i^{\text{in}} \int_{-\infty}^{\infty} \epsilon(t)\,\mathrm{d}t}_{=:c^{\text{out}}} + \sum_{i=1}^{n} w_i \left[\epsilon \circ a_i^{\text{in}}\right](t) \tag{7.42}$$

$$= c^{\text{out}} + \left[\epsilon \circ \sum_{i=1}^{n} w_i\, a_i^{\text{in}}\right](t) \tag{7.43}$$

$$\overset{(7.1)}{=} c^{\text{out}} + \left[\epsilon \circ a^{\text{out}}\right](t)\,, \tag{7.44}$$

and the joint firing rate is (Kempter et al., 1999)

$$r_i^{\text{in,out}}(t,t') := \langle S_i^{\text{in}}(t) S^{\text{out}}(t') \rangle_{E^{\text{in}},E^{\text{out}}} \tag{7.45}$$

$$= r_i^{\text{in}}(t) r^{\text{out}}(t') + w_i \epsilon(t'-t) r_i^{\text{in}}(t). \tag{7.46}$$

The first term would result also from a rate model, while the second term captures the statistical dependencies between input and output spike-trains mediated by the synaptic weights w_i and the EPSP ϵ.

7.3.2 STDP Can Perform SFA

In this section we will demonstrate that in an ensemble-averaged sense it is possible to generate the same weight distribution as in the continuous model by means of an STDP rule with a specific learning window.

Synaptic plasticity that depends on the temporal order of pre- and postsynaptic spikes has been found in a number of neuronal systems (Debanne et al., 1994; Markram et al., 1997; Bi & Poo, 1998; Zhang et al., 1998; Feldman, 2000), and has raised a lot of interest among modelers (Gerstner et al., 1996; Abbott & Blum, 1996; for a review see Kepecs et al., 2002). Typically, synapses undergo long-term potentiation (LTP) if a presynaptic spike precedes a postsynaptic spike within a time scale of tens of milliseconds and long-term depression (LTD) for the opposite temporal order. Assuming that the change in synaptic efficacy occurs on a slower timescale than the typical interspike interval, the STDP weight dynamics can be modeled as

$$\Delta w_i = \eta \sum_{\alpha}^{m_i^{\text{in}}} \sum_{\beta}^{m^{\text{out}}} W(t_{i\alpha}^{\text{in}} - t_{\beta}^{\text{out}})\,. \tag{7.47}$$

Here $t_{i\alpha}^{\text{in}}$ denotes the spike times of the presynaptic spikes at synapse i and t_{β}^{out} denotes the postsynaptic spike times. $W(t)$ is the learning window that determines if and to what extent the synapse is potentiated or depressed by a single spike pair. The convention

is such that negative arguments t in $W(t)$ correspond to the situation where the presynaptic spike precedes the postsynaptic spike. m^{in} and m^{out} are the numbers of pre- and postsynaptic spikes occurring in the time interval $[t_a, t_b]$ under consideration. η is a small positive learning rate. Note that due to the presence of this learning rate the absolute scale of the learning window W is not important for our analysis.

We circumvent the well-known stability problem of STDP by applying an explicit weight normalization ($\vec{w}^{\text{new}} = (\vec{w}^{\text{old}} + \Delta\vec{w})/||\vec{w}^{\text{old}} + \Delta\vec{w}||$) instead of weight-dependent learning rates as used elsewhere (Kistler & Hemmen, 2000; Rubin et al., 2001; Gütig et al., 2003). Such a normalization procedure could be implemented by means of a homeostatic mechanism targeting the output firing rate (e.g., by synaptic scaling; for reviews see (Turrigiano & Nelson, 2000; Abbott & Nelson, 2000)).

Modeling the spike trains as sums of delta pulses (i.e., $S^{\text{in/out}} = \sum_j \delta(t - t_j^{\text{in/out}})$) the learning rule (7.47) can be rewritten as

$$\Delta w_i = \eta \int_{t_a}^{t_b}\int_{t_a}^{t_b} W(t-t')S_i^{\text{in}}(t)S^{\text{out}}(t')\,\mathrm{d}t\,\mathrm{d}t' \tag{7.48}$$

$$\approx \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')\hat{S}_i^{\text{in}}(t)\hat{S}^{\text{out}}(t')\,\mathrm{d}t\,\mathrm{d}t'. \tag{7.49}$$

Taking the ensemble average allows us to retrieve the rates that underlie the spike trains and thus the signals $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$ of the continuous model:

$$\langle \Delta w_i \rangle_{E^{\text{in}},E^{\text{out}}} \overset{(7.49)}{\approx} \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')\langle \hat{S}_i^{\text{in}}(t)\hat{S}^{\text{out}}(t')\rangle_{E^{\text{in}},E^{\text{out}}}\,\mathrm{d}t\,\mathrm{d}t' \tag{7.50}$$

$$\overset{(7.46)}{=} \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')\big(\hat{r}_i^{\text{in}}(t)\hat{r}^{\text{out}}(t') + w_i\epsilon(t'-t)\hat{r}_i^{\text{in}}(t)\big)\,\mathrm{d}t\,\mathrm{d}t' \tag{7.51}$$

$$\overset{(7.36,7.44)}{\approx} \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')[c_i^{\text{in}} + \hat{a}_i^{\text{in}}](t)[c^{\text{out}} + \epsilon \circ \hat{a}^{\text{out}}](t')\,\mathrm{d}t\,\mathrm{d}t'$$

$$+ \quad \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')w_i\epsilon(t'-t)[c_i^{\text{in}} + \hat{a}_i^{\text{in}}](t)\,\mathrm{d}t\,\mathrm{d}t'\,. \tag{7.52}$$

Expanding the products in equation (7.52) gives rise to a number of terms, among which only one depends on both the input and the output signal $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$. Because each input signal has vanishing mean, terms containing just one input signal lead to negligible contributions. The remaining terms depend only on the mean firing rates c_i^{in} and c^{out}:

$$\begin{aligned}\langle \Delta w_i \rangle_{E^{\text{in}},E^{\text{out}}} &\overset{(7.52)}{\approx} \eta \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(t-t')\hat{a}_i^{\text{in}}(t)[\epsilon \circ \hat{a}^{\text{out}}](t')\,\mathrm{d}t\,\mathrm{d}t' \\ &+ \quad \eta w_i c_i^{\text{in}} T_{ab} \int_{-\infty}^{\infty} W(t)\epsilon(-t)\,\mathrm{d}t \qquad (7.53) \\ &+ \quad \eta c^{\text{out}} c_i^{\text{in}} T_{ab} \int_{-\infty}^{\infty} W(t)\,\mathrm{d}t\,.\end{aligned}$$

A generalized version of equation (7.53) that incorporates non-Hebbian plasticity (i.e., terms that depend on the pre/postsynaptic signals only) has been derived and discussed by Kempter et al. (2001). Regarding the effects of the input signals on learning, the decisive term is the first one. The other two are rather unspecific in that they do not depend on the properties of the input and output signals $\hat{a}_i^{\text{in}}$ and $\hat{a}^{\text{out}}$.

The second term alone would generate a competition between the weights: synapses that experience a higher mean input firing rate c_i^{in} grow more rapidly than those with smaller input firing rates. If we assume that the input neurons fire with the same mean firing rate, all weights grow with the same rate, so the direction of the weight vector remains unchanged. Thus, due to the explicit weight normalization this term has no effect on the weight dynamics and can be neglected.

If the integral over the learning window is positive, the third term in equation (7.53) favors a weight vector that is proportional to the vector of the mean firing rates of the input neurons. It thus stabilizes the homogeneous weight distribution and opposes the effect of the first term, which captures correlations in the input signals. Note that this is only true if the integral over the learning window is positive: otherwise, this term introduces a competition between the weights (Kempter et al., 2001; Gütig et al., 2003). One possible interpretation is that the neuron has a "default state" in which all synapses are equally strong and that correlations in the input need to surpass a certain threshold in order to be imprinted in the synaptic connections. Interestingly, this threshold is determined by the integral over the learning window, which implies that neurons that balance LTP and LTD should be more sensitive to input correlations.

An alternative possibility is that the neuron possesses a mechanism of canceling the effects of this term. From a computational perspective this would be sensible, as the mean firing rates c_i^{in} and c^{out} do not carry information about the input, neither in rate nor in a timing code. If we conceive neurons as information encoders aiming at adapting to the structure of their input, this term is thus more hindrance than help. Assuming that the neuron compensates for this term, the dynamics of the synaptic weights are governed exclusively by the correlations in the input signals as reflected by the first term. In the following we will restrict our considerations to this term and omit the others.

Rearranging the temporal integrations, we can rewrite equation (7.53) for the weight updates as

$$\langle \Delta w_i \rangle_{E^{\text{in}},E^{\text{out}}} \overset{(7.53)}{\approx} \eta \int_{-\infty}^{\infty} [W \circ \epsilon](t)[\hat{a}^{\text{out}} \star \hat{a}_i^{\text{in}}](t)\,\mathrm{d}t\,. \qquad (7.54)$$

The first conclusion we can draw from this reformulation is that for the dynamics of the learning process the convolution of the learning window with the EPSP and not

the learning window alone is relevant. As discussed below, this might have important consequences for functional interpretations of the shape of the learning window.

Second, by comparison with equation (7.29), it is obvious that in order to learn the same weight distribution as in the continuous model, the learning window has to fulfill the condition that

$$[W \circ \epsilon](t) = [f^{\text{in}} \star f^{\text{out}}](t) \;=:\; W_0(t) \tag{7.55}$$

$$\iff \quad \mathcal{F}_{W\circ\epsilon}(\nu) \;=\; \mathcal{F}_W(\nu)\mathcal{F}_\epsilon(\nu) = \mathcal{F}_{f^{\text{in}}\star f^{\text{out}}}(\nu) \;=\; \mathcal{P}_{f_{\text{SFA}}}(\nu) \;=\; \mathcal{F}_{W_0}(\nu). \tag{7.56}$$

Here, W_0 is the convolution of W with ϵ and is equal to the learning window in the limit of an infinitely short, δ-shaped EPSP. As the power spectrum $\mathcal{P}_{f_{\text{SFA}}}(\nu)$ is of course real, W_0 is symmetric in time. Note that the width of W_0 scales inversely with the width of the power spectrum $\mathcal{P}_{f_{\text{SFA}}}$, which in turn is proportional to ν_{max}. Once the power spectrum $\mathcal{P}_{f_{\text{SFA}}}$ and the EPSP is given, equation (7.56) uniquely determines the learning window W. Because it is W_0 rather than W that determines the learning dynamics, we will refer to W_0 as the "effective learning window".

7.3.3 Learning Windows

According to the last section, we require special learning windows in order to learn the slow directions in the input. This of course raises the question of which window shapes are favorable, and in particular if these are in agreement with physiological findings.

Given the shape of the EPSP and the power spectrum $\mathcal{P}_{f_{\text{SFA}}}$, the learning window is uniquely determined by equation (7.56). Remember that the only parameter in the power spectrum $\mathcal{P}_{f_{\text{SFA}}}$ is the frequency ν_{max}, above which the power spectrum of the input data was assumed to vanish. For simplicity, we model the EPSP as a single exponential with a time constant τ:

$$\epsilon(t) = \theta(t)\, \mathrm{e}^{-\frac{t}{\tau}}\,. \tag{7.57}$$

For this particular EPSP shape, the learning window can be calculated analytically by inverting the Fourier transform in equation (7.56). The result can be written as

$$W(t) = \left[\frac{\mathrm{d}}{\mathrm{d}t} + \frac{1}{\tau}\right] W_0(t)\,. \tag{7.58}$$

W_0 is symmetric, so its derivative is antisymmetric. Thus, the learning window is a linear combination of a symmetric and an antisymmetric component. As the width of W_0 scales with the inverse of ν_{max}, its temporal derivative scales with ν_{max}. Accordingly, the symmetry of the learning window is governed by an interplay of the duration τ of the EPSP and the maximal input frequency ν_{max}. For $\tau \ll 1/\nu_{max}$ the learning window is dominated by W_0 and thus symmetric whereas for $\tau \gg 1/\nu_{max}$ the temporal derivative of W_0 is dominant, so the learning window is antisymmetric.

We have assumed that the input signals have negligible power above the maximal input frequency ν_{max}. Thus, the temporal structure of the input signals can only provide a lower bound for ν_{max}. On the other hand, exceedingly high values for ν_{max} lead to very narrow learning windows, thereby sharpening the coincidence detection and reducing the speed of learning. Moreover, it may be metabolically costly to implement physiological processes that are faster than necessary. Thus, it appears sensible to choose ν_{max} such that $1/\nu_{max}$ reflects the fastest timescale in the input signals. Accordingly, the symmetry of the learning window is governed by the relation between the length of the EPSP and

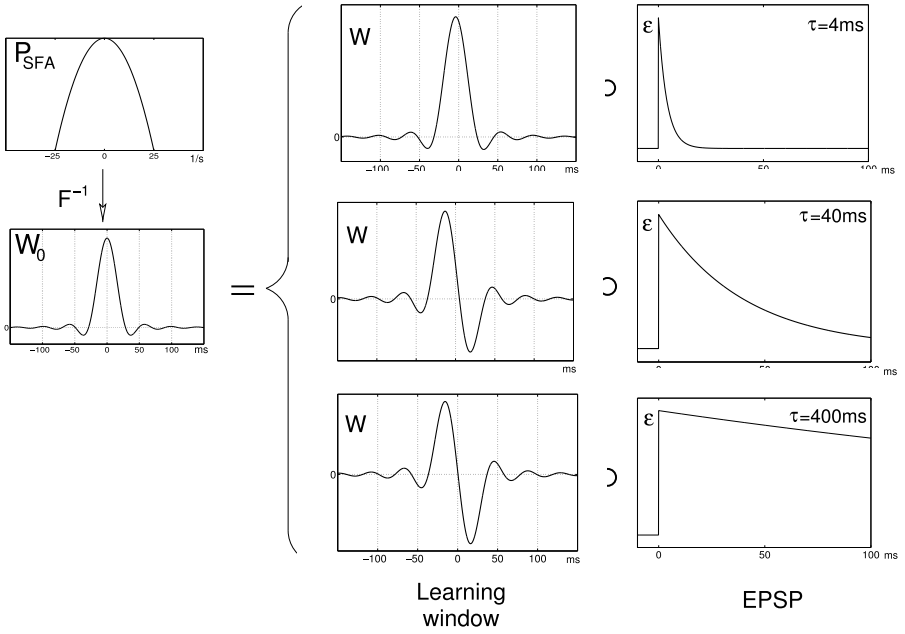


Figure 7.3: **Relation between the EPSP and the learning window**. The power spectrum $\mathcal{P}_{f_{SFA}}$ is the Fourier transform of the effective learning window W_0, which in turn is the convolution of the learning window W and the EPSP ϵ. The figure shows the learning windows required for SFA for three different EPSP durations ($\tau = 4, 40, 400$ms). The maximal input frequency ν_{max} was $1/(40\text{ms})$ in all plots.

the fastest time scale in the input data. If the EPSP is short enough to resolve the fastest input components, the learning window is symmetric. If the EPSP is too long to fully resolve the temporal structure of the input (i.e., it acts as a low-pass filter), the learning window will tend to be antisymmetric.

We choose a value of $\nu_{max} = 1/(40\text{ms})$. The argument for this choice is that within a rate code, the cells that project to the neuron under consideration can hardly convey signals that vary on a faster time scale than the duration of their EPSP. It is thus reasonable to choose the time constant of the EPSP and the inverse of the cutoff frequency to have the same order of magnitude. Typical durations of cortical EPSPs are of the order of tens of milliseconds (see Koch et al. (1996) for further references and a critical discussion), so 40 ms seems a reasonable value.

Figure 7.3 illustrates the connection between $\mathcal{P}_{f_{SFA}}$, W_0, the learning window, and the EPSP. It also shows the learning windows for three different durations of the EPSP, while keeping ν_{max} =1/(40ms). The oscillatory and slowly decaying tails of $W(t)$ are due to the sharp cutoff of the power spectrum $\mathcal{P}_{f_{SFA}}$ at $|\nu| = \nu_{max}$ and become less pronounced if $\mathcal{P}_{f_{SFA}}$ is smoothened out.

As negative time arguments in $W(t)$ correspond to the case in which the presynaptic spike (and thus the onset of the resulting EPSP) precedes the postsynaptic spike, the shape of the theoretically derived learning window for physiologically plausible values of τ and ν_{max} ($\tau = 1/\nu_{max} = 40$ms, middle row in figure 7.3) predicts potentiation of the synapse when a postsynaptic spike is preceded by the onset of an EPSP and depression of the synapse when this temporal order is reversed. This behavior is in agreement with experimental data from neocortex and hippocampus in rats as well as from the optic

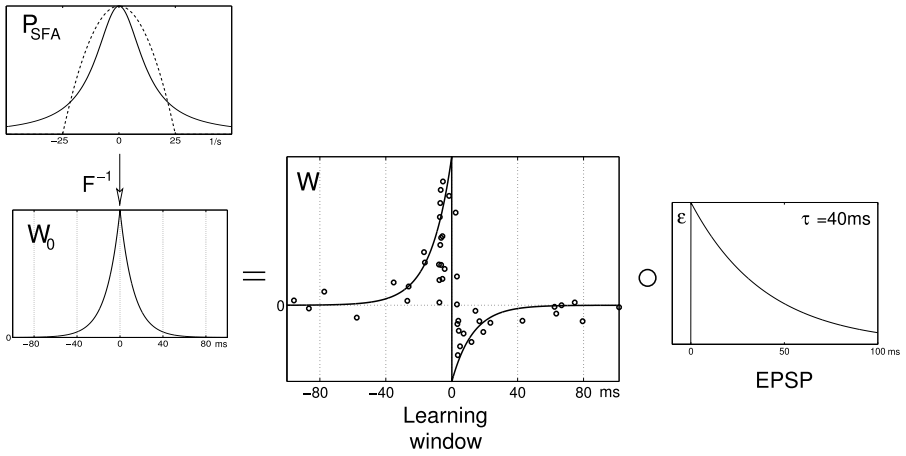


Figure 7.4: **Comparison of the learning window with experimental data**. The plot compares the theoretically predicted learning window with experimental data from hippocampal pyramidal cells as published by Bi & Poo (1998) (larger plot in the middle). Instead of the ideal power spectrum $\mathcal{P}_{f_{\mathrm{SFA}}}$ with the abrupt cutoff at ν_{max} as stated in equation (7.20) a Cauchy function with γ=1/(15ms) was used (top left, the dashed line is $\mathcal{P}_{f_{\mathrm{SFA}}}$ for ν_{max}=1/(40ms)). Again, the EPSP decay time was τ = 40ms. This learning window corresponds to an implementation of the 'trace rule' (Földiàk, 1991; O'Reilly & Johnson, 1994; Wallis & Rolls, 1997) for a decay time of the exponential filter of 15ms.

tectum in *Xenopus* (Debanne et al., 1994; Bi & Poo, 1998; Feldman, 2000; Markram et al., 1997; Zhang et al., 1998). To further illustrate this agreement, Figure 7.4 compares the data as published by Bi & Poo (1998) with the learning window resulting from a smoothened power spectrum with the shape of a Cauchy function (7.35) instead of $\mathcal{P}_{f_{\mathrm{SFA}}}$. As demonstrated above, this corresponds to implementing the slowness principle in form of the trace rule. Interestingly, the resulting learning window has the double-exponential shape that is regularly used in models of STDP (e.g., in Rossum et al. (2000); Song & Abbott (2001); Gütig et al. (2003)). As the absolute scale of the learning window is not determined in our analysis, it was adjusted to facilitate the comparison with the experimental data.

7.3.4 Interpretation of the Learning Windows

The last section leaves a central question open: why are these learning windows optimal for slowness learning and why does the EPSP play such an important role for the shape of the learning window?

Let us first discuss the case of the symmetric learning window, that is the situation in which the EPSP is shorter than the fastest time scale in the input signal. Then, the convolution with the EPSP has practically no effect on the temporal structure of the signal and the output firing rate can be regarded as an instantaneous function of the input rates. We can thus neglect the EPSP altogether. The learning mechanism can then be understood as follows: assume at a given time t the postsynaptic firing rate r^{out} is high and causes a postsynaptic spike. Then the finite width of the learning window leads to potentiation not only of those synapses that participated in initiating the spike but also of those that transmit a spike within a certain time window around the time of the postsynaptic spike. As this leads to an increase of the firing rate within this time window, the learning mechanism tends to equilibrate the firing rates for neighboring times and

thus favors temporally slow output signals.

If the duration of the EPSP is longer than the fastest time scale in the input signal, the output firing rate is no longer an instantaneous function of the input signals but generated by low-pass filtering the signal a^{out} with the EPSP. This affects learning, because the objective of the continuous model is to optimize the slowness of a^{out}, whose temporal structure is now "obscured" by the EPSP. In order to optimize the objective, the system thus has to develop a deconvolution mechanism to reconstruct a^{out}. From this point of view, the learning window has to perform two tasks simultaneously. It has to first perform the deconvolution and then enforce slowness on the resulting signal. This is most easily illustrated by means of the condition (7.55). The convolution of the learning window with the EPSP generates the effective learning window $W_0(t)$ that is independent of the EPSP and which coincides with the learning window for infinitely short EPSPs. Intuitively, we could solve equation (7.55) by choosing a learning window that consists of the "inverse" of the EPSP and the EPSP-free learning window W_0. An intuitive example is the limiting case of an infinitely long EPSP. The EPSP then corresponds to a Heaviside function and performs an integration, which can be inverted by taking the derivative. Thus, the learning window for long EPSPs is the temporal derivative of the learning window for short EPSPs. The dependence of the required learning window on the shape of the EPSP is thus caused by the need of the learning window to "invert" the EPSP.

These considerations shed a different light on the shape of physiologically measured learning windows. The antisymmetry of the learning window may not act as a physiological implementation of a causality detector after all, but rather as a mechanism for compensating intrinsic low-pass filters in neuronal processing such as the EPSP. For functional interpretations of STDP, it may be more sensible to consider the convolution of the learning window with the EPSP than the learning window alone.

It should be noted that, according to our learning rule, the weights adapt in order to make a hypothetical instantaneous output signal a^{out} optimally slow. This does not necessarily imply that the output firing rate r^{out}, which is generated by low-pass filtering a^{out} with the EPSP, is optimally slow. In principle, the system could generate more slowly varying signals by exploiting the temporal structure of the EPSP. However, the motivation for the slowness principle is the idea that the system learns to detect invariances in the *input* signal, and that from this perspective the goal of creating a slowly varying output signal is not an end in itself but a means to learn invariances. Thus, the low-pass filtering effect of the EPSP should not be exploited but ignored or compensated.

7.3.5 General Learning Windows and EPSPs

Although the asymmetry in LTP/LTD induction observed by Bi & Poo (1998) has also been observed in other studies, the decay times for the LTP and the LTD branches of the learning window appear to be different in other preparations (Feldman, 2000). One may thus ask how robust our interpretation is with respect to the detailed shape of the learning window. To address this question, we start with some general learning window W and EPSP ϵ and ask under which conditions the effective learning window $W_0 = W \circ \epsilon$ prefers slowly varying features in the input.

As a starting point we use the dynamics (7.54) of the weights as generated by the input statistics. Using $a^{\text{out}} = \sum_j w_j a_j^{\text{in}}$ and defining the correlation functions

$$C_{ij}(t) := [a_j^{\text{in}} \star a_i^{\text{in}}](t) \tag{7.59}$$

yields

$$\langle \Delta w_i \rangle_{E^{\text{in}},E^{\text{out}}} = \sum_j \underbrace{\left[\eta \int W_0(t) C_{ij}(t) \mathrm{d}t\right]}_{=:A_{ij}} w_j \,. \tag{7.60}$$

The dynamics thus follows a linear difference equation with a dynamic matrix A_{ij} whose properties are determined by the correlation function $C_{ij}(t)$ and the effective learning window $W_0(t)$. One important question is whether the weights approach a stable fixed-point state or oscillate. In this context, the symmetry properties of A_{ij} and thus those of C_{ij} are crucial. The correlation functions obey the relation

$$C_{ij}(t) = C_{ji}(-t) \,, \tag{7.61}$$

which couples their spatial symmetry (i.e., the symmetry with respect to the indices i and j) to their temporal symmetry. For instance, if the input statistics is reversible, i.e., for $C_{ij}(t) = C_{ij}(-t)$, C_{ij} is symmetric in the indices and so is A_{ij}. If the input statistics were 'perfectly irreversible', i.e., $C_{ij}(-t) = -C_{ij}(t)$, C_{ij} and A_{ij} would be antisymmetric. This motivates the splitting of the correlation functions C_{ij} into a temporally symmetric and an antisymmetric component: $C_{ij} = C^+_{ij} + C^-_{ij}$ with $C^\pm_{ij}(t) = \pm C^\pm_{ij}(-t)$. In a similar fashion we split the effective learning window $W_0(t) = W_0^+(t) + W_0^-(t)$. For symmetry reasons, the dynamical matrix A_{ij} can then be separated into two components

$$A_{ij} = \underbrace{\eta \int W_0^+(t) C^+_{ij}(t) \mathrm{d}t}_{=:A^+_{ij}} + \underbrace{\eta \int W_0^-(t) C^-_{ij}(t) \mathrm{d}t}_{=:A^-_{ij}} \,. \tag{7.62}$$

Because of the symmetry relation (7.61), A^+_{ij} is symmetric in i and j, while A^-_{ij} is antisymmetric.

This shows that the effective learning window W_0 can be split into two functionally different components. The symmetric component picks up the reversible aspects of the input statistics while the antisymmetric component detects irreversibilities, e.g., possible causal relations within the input data. It is this antisymmetric component of the learning window that has previously been interpreted as a means for sequence learning and predictive coding (Abbott & Blum, 1996; Rao & Sejnowski, 2001). Note that the associated weight update $\sum_j A^-_{ij} w_j$ is always orthogonal to the weight itself. Thus, irreversibilities in the input data in combination with an antisymmetric learning window work against the development of a stable weight distribution, even if the input statistics is stationary. In particular, weight oscillations on the time scale of learning may occur. For instance, in networks with recurrent connections that learn according to STDP, previous studies have shown that the network tends to develop a state of distributed synchrony (Horn et al., 2000) that resembles synfire chains. These activity patterns display a pronounced causal structure, so it would be interesting to check if the synaptic weights that emerge in such a network are stable or show oscillations. It is likely that in this context the model constraints on the weights play an important role. If the weights are limited by hard boundaries (Horn et al., 2000), they tend to saturate, thereby avoiding oscillatory solutions. In the case of softer weight constraints, e.g., in models of STDP with multiplicative weight-dependence, oscillations may occur.

If W_0 is symmetric or if the input statistics is reversible, $C^-_{ij} = 0$, the dynamical matrix $A_{ij} = A^+_{ij}$ is symmetric. As already seen for the case of the continuous model neuron, the

learning dynamics can then be interpreted as a gradient ascent on the objective function

$$\Psi = \frac{1}{2}\sum_{i,j} w_i A^+_{ij} w_j = \frac{1}{2}\int W_0^+(\nu)\mathcal{P}_{a^{\text{out}}}(\nu)\mathrm{d}\nu\,. \tag{7.63}$$

As discussed earlier, this objective function can be interpreted as an implementation of the slowness principle if $W_0^+(\nu)$ is a low-pass filter, i.e., if it has a global maximum at zero frequency. This indicates that at least for reversible input statistics the preference of STDP for slow signals may be rather insensitive to details of the learning window.

7.4 Discussion

As discussed in the introduction, the algorithm that underlies SFA is rather technical. Here, we have examined whether it is feasible to implement SFA within the limitations of neuronal circuitry. We have approached this question analytically and demonstrated that such an implementation is possible in both continuous and spiking model neurons.

In the first part of the chapter, we have shown that for linear continuous model neurons, the slowest direction in the input signal can be learned by means of Hebbian learning on low-pass filtered versions of the input and the output signal. The power spectrum of the low-pass filter required for implementing SFA can be derived from the learning objective and has the shape of an upside-down parabola.

The idea of using low-pass filtered signals for invariance learning is a feature that our model has in common with several others (Földiàk, 1991; O'Reilly & Johnson, 1994; Wallis & Rolls, 1997). By means of the continuous model, we have discussed the relation of our model to these 'trace rules' and have shown that they bear strong similarities.

In the second part of the chapter we have discussed the modifications that have to be made to adjust the learning rule for a Poisson neuron. We find that in an ensemble-averaged sense it is possible to reproduce the behavior of the continuous model neuron by means of spike-timing–dependent plasticity (STDP). Moreover, the analysis suggests that the outcome of STDP learning is not governed by the learning window alone but rather by the convolution of the learning window with the EPSP, which is of relevance for functional interpretations of STDP.

The learning window that realizes SFA can be calculated analytically. Its shape is determined by the interplay of the duration of the EPSP and the maximal input frequency ν_{max}, above which the input signals are assumed to have negligible power. If ν_{max} is small, i.e., if the EPSP is sufficiently short to temporally resolve the most quickly varying components of the input data, the learning window is symmetric whereas for large ν_{max} or long EPSPs, it is antisymmetric. Interestingly, physiologically plausible parameters lead to a learning window whose shape and width is in agreement with experimental findings. Based on this result, we propose a new functional interpretation of the STDP learning window as an implementation of the slowness principle that compensates for neuronal low-pass filters such as the EPSP.

An important question in this context is on which timescales is this interpretation valid. It is conceivable that for signals that vary on a time scale of less than a hundred milliseconds, a learning window with a width of tens of milliseconds can distinguish slower from faster signals. STDP could thus be sufficient to establish invariant representations in early sensory processing, e.g. visual receptive fields that become invariant to microsaccades inducing small translations. Although it is unlikely that STDP alone can

distinguish between signals that vary on behavioral time scales of hundreds of milliseconds or even seconds, this may not be problematic, because it is probably not sensible to order *all* aspects of the stimuli according to how quickly they vary. Rather, one should distinguish input components that vary so quickly that they are unlikely to be behaviorally relevant from those that vary on behavioral time scales. From this perspective the intrinsic time scale of the learning rule should be such that its discriminative power is best on a time scale where this transition occurs. It is conceivable that this transition time scale is on the order of several tens of milliseconds. The learning of high level invariances that correspond to behavioral time scales will probably require additional mechanisms with corresponding intrinsic time scales, e.g., sustained firing in response to a stimulus (Drew & Abbott, 2006).

For general learning windows and EPSPs, the convolution of the learning window with the EPSP can be split into a symmetric and an anti-symmetric component. The symmetric component picks up reversible aspects of the input statistics while the anti-symmetric component detects irreversible aspects. Previous functional interpretations of STDP have mostly concentrated on the antisymmetric component, which has been interpreted, e.g., as a mechanism for sequence learning or predictive coding (Rao & Sejnowski, 2001; Abbott & Blum, 1996) or for reducing recurrent connectivity in favor of feed-forward structures (Horn et al., 2000; Song & Abbott, 2001). Other studies have neglected the phase structure of the learning window altogether and concentrated on its power spectrum, proposing that timing-dependent plasticity performs Hebbian learning on an optimal estimate of the input signals in the presence of noise (Wallis & Baddeley, 1997; Dayan et al., 2004). Note that these interpretations are not necessarily contradictory to ours, because the slowness interpretation relies on the symmetric component of the learning window only and thus on the reversible aspect of the input statistics. These considerations indicate that depending on the temporal structure of the input, STDP may have different functional roles.

A different approach to unsupervised learning of invariances with a biologically realistic model neuron has been taken by K. P. Körding & König (2001). In their model, bursts of backpropagating spikes gate synaptic plasticity by providing sufficient amounts of dendritic depolarization. These bursts are assumed to be triggered by lateral connections that evoke calcium spikes in the apical dendrites of cortical pyramidal cells.

Of course the model presented here is not a complete implementation of SFA. We have only considered the central step of SFA, the extraction of the most slowly varying direction from a set of whitened input signals. To implement the full algorithm, additional steps are necessary: a nonlinear expansion of the input space, the whitening of the expanded input signals and a means of normalizing the weights. When traversing the dendritic arborizations of a postsynaptic neuron, axons often make more than one synaptic contact. As different input channels may be subjected to different nonlinearities in the dendritic tree (cf. London & Häusser, 2005), the postsynaptic neuron may have access to several nonlinearly transformed versions of the same presynaptic signals. Conceptually, this resembles a nonlinear expansion of the input signals. However, it is not obvious, how these signals could be whitened within the dendrite. On the network level, however, whitening could be achieved by adaptive recurrent inhibition between the neurons (Földiàk, 1989). This mechanism may also be suitable for extracting several slow uncorrelated signals as required in the original formulation of SFA (Wiskott & Sejnowski, 2002) instead of just one. We assumed an explicit weight normalization in the description

of our model. Alternatively, one could also use a modified learning rule that implicitly normalizes the weight vector as long as it extracts the signal with the largest variance. A possible biological mechanism is synaptic scaling (Turrigiano & Nelson, 2000), which is believed to multiplicatively rescale all synaptic weights according to postsynaptic activity, similar to Oja's rule (Oja, 1982; Abbott & Nelson, 2000). Thus, it appears that most of the mechanisms necessary for an implementation of the full SFA algorithm are available, but that it is not yet clear how to combine them in a biologically plausible way.

Another critical point in the analytical derivation for the spiking model is the replacement of the temporal by the ensemble average, as this allows recovery of the rates that underlie the Poisson processes. The validity of the analytical results thus requires some kind of ergodicity in the training data, a condition which of course needs to be justified for the specific input data at hand.

It is still open if the results presented here can be reproduced with more realistic model neurons. The spiking model neuron used here was simplified in that it had a linear relationship between input and output firing rate. In many real neurons, highly nonlinear behavior was observed. Interestingly, Hebbian learning for nonlinear rate-based neurons has previously been associated with the detection of higher-order moments of the input statistics (Oja & Karhunen, 1995), thereby providing a mechanism for extracting statistically independent components of the input signal. Because for sparse input statistics independent component analysis is closely related to sparse coding (Olshausen & Field, 1997), it is tempting to speculate that within a rate picture temporally nonlocal plasticity with a nonlinear input-output relation implements a combination of sparseness and slowness. Learning paradigms that combine these two objectives are thus an interesting field for further studies (Franzius, Sprekeler & Wiskott, 2007; Blaschke et al., 2007).

Another nonlinearity that we have neglected is the frequency- and weight-dependence of STDP (Bi & Poo, 1998; Sjöström et al., 2001). Additional work will be needed to examine how these interfere with the proposed functional role of STDP. Furthermore, modeling the spiking mechanism of a neuron by an inhomogeneous Poisson process is also a severe simplification that ignores basic phenomena of spike generation in biological neurons such as refractoriness and thresholding. It is not clear how these characteristics would change the learning rule that leads to an implementation of the slowness principle. It seems to be a very difficult task to answer these questions analytically. Simulations will be necessary to verify the results derived here and to analyze which changes appear and which adaptations must be made in a more realistic model of neural information processing.

In summary, the analytical considerations presented in this chapter show that (i) slowness can be equivalently achieved by minimizing the variance of the time derivative signal or by maximizing the variance of the low-pass filtered signal, the latter of which can be achieved by standard Hebbian learning on the low-pass filtered input and output signals; (ii) the difference between SFA and the trace learning rule lies in the exact shape of the effective low-pass filter - for most practical purposes the results are probably equivalent; (iii) for a spiking Poisson model neuron with an STDP learning rule, it is not the learning window that governs the weight dynamics but the convolution of the learning window with the EPSP; (iv) the STDP learning window that implements the slowness objective is in good agreement with learning windows found experimentally. With these results, we have reduced the gap between slowness as an abstract learning principle and biologi-

cally plausible STDP learning rules and we offer a completely new interpretation of the standard STDP learning window.

Chapter 8

Outlook: Towards Reaction-Diffusion Systems

8.1 Introduction

As discussed in the last chapter, the implementation of SFA as a batch learning algorithm is problematic from the viewpoint of biological plausibility, so that an implementation of the slowness principle in terms of an online learning rule would be more favorable. We have also shown that under certain conditions, temporally nonlocal Hebbian plasticity can be interpreted as a gradient descent on an objective function that includes slowness. This suggests further analysis of gradient-based slowness learning.

In contrast to batch learning algorithms like SFA, online learning rules give access to the dynamics of the learning process. Moreover, because the internal parameters (e.g., the synaptic weights) directly determine the response behavior or receptive field (RF) of the neuron, the learning dynamics implicitly determines the dynamics of the neuron's RF. In this chapter, we will outline how this dependence can be made explicit by deriving dynamics equations for the RF. We will approach this problem from two directions: From the level of slowness as an abstract optimization principle and from level of a simple model for synaptic plasticity. Schematically, these approaches is illustrated in figure 8.1. The first approach will show that gradient descent on the Δ-value leads to a diffusion equation for the RF. In the second approach, we will further analyze the learning dynamics of STDP and show that under certain assumptions, the dynamics of the RF under STDP obey a drift-diffusion equation. Thus, both approaches lead to the same class of dynamical equations for the receptive field. The diffusion components in the dynamics can be interpreted as an implementation of invariance learning.

Furthermore, we will discuss the effects of constraints within the gradient-based approach. We show that the introduction of punishment terms for unit variance and decorrelation in the objective function leads to additional contributions in the RF dynamics that can be interpreted in terms of homeostasic plasticity and lateral interactions with anti-Hebbian learning. The resulting equations have the structure of a reaction-diffusion system. This class of systems is somewhat canonical in the wide field of self-organization. Our hope is that the analysis provides a starting point for a bridge between rules of synaptic plasticity and the analysis of receptive field dynamics in terms of established methods of nonlinear dynamics.

This chapter is conceptual in nature and meant to provide an outlook on two math-

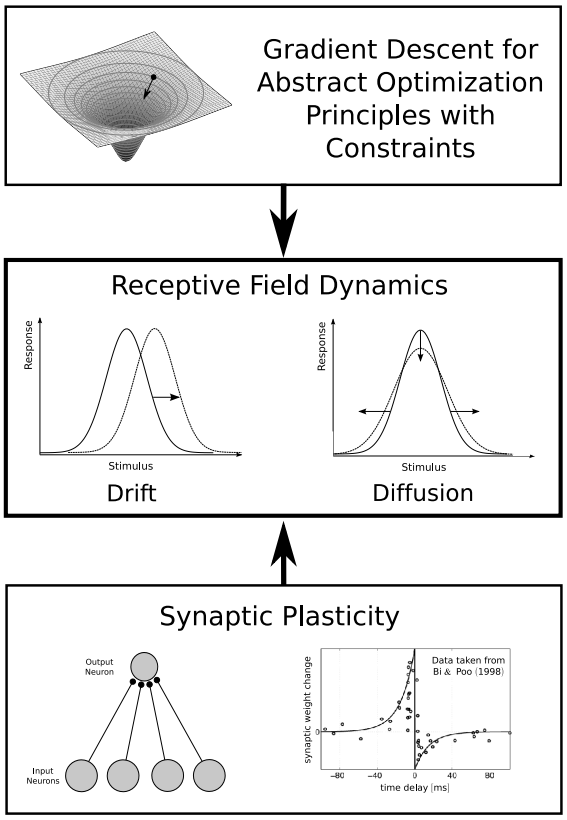


Figure 8.1: **Schema of the approaches in this chapter.** Receptive field dynamics are derived from two different perspectives. Starting with the abstract optimization problem of SFA (topmost box), we use gradient descent to derive dynamical equations for receptive fields. Secondly, we start with a phenomenological model of physiologically characterized synaptic plasticity (bottommost box). Together with models for the presynaptic receptive fields and for the processing behavior of the output neuron, the learning dynamics of the synaptic weights also lead to dynamical equations for the receptive field of the output neuron. For both approaches, the receptive field dynamics contain drift and diffusion dynamics of the receptive fields (center box).

ematical approaches to receptive field formation. The formalism presented here is yet to be applied to concrete problems.

8.2 Receptive Field Dynamics of Uncoupled Neurons

8.2.1 Gradient-Based Slowness Learning

Assumptions

Like in chapter 3, we assume that we have access to an unrestricted function space from which we choose functions $g_j(\mathbf{x})$ of the input data $\mathbf{x}$ that generate output signals $y_j(t) = g_j(\mathbf{x}(t))$. For notational simplicity, we arrange the functions g_j in a vector $\mathbf{g} = (g_1, ..., g_J)$. Again, the input data are described in terms of probability distributions $p_{\mathbf{x},\dot{\mathbf{x}}}$.

Gradient Descent on Functionals

The standard approach for deriving an online learning rule that minimizes a given objective function is gradient descent. Say we want to minimize an objective function $\Phi(\mathbf{w})$ that depends on some internal parameters $\mathbf{w}$. Then, gradient descent means that we change the parameters in the direction where the objective function declines most steeply, i.e., in the opposite direction of the gradient:

$$\partial_\tau \mathbf{w} = -\eta \nabla_{\mathbf{w}} \Phi(\mathbf{w}), \tag{8.1}$$

where τ is a coarse-grained time scale, on which learning takes place, η is a learning rate, and $\nabla_{\mathbf{w}}\Phi(\mathbf{w})$ denotes the gradient of the objective function with respect to the internal parameters $\mathbf{w}$.

This formulation of gradient descent is valid when there is a finite number of internal parameters, i.e., where $\mathbf{w}$ is finite-dimensional. Since we assume that the functions $\mathbf{g}$ are unrestricted, any objective function $\Phi[\mathbf{g}]$ must necessarily be a *functional* of the functions $\mathbf{g}$, so the number of internal parameters is infinite. Thus, we need to calculate a "gradient" in an infinite-dimensional function space $\mathcal{F}$. Clearly, the finite-dimensional concept of a gradient has to be modified in such a scenario.

Geometrically, the gradient points in the direction in which an infinitesimal perturbation of fixed length leads to the maximal change in the objective. This concept can be generalized to infinite-dimensional scenarios by means of variational calculus. For illustration, let us first consider an objective function $\Psi[f]$ that depends on a single scalar function $f(\mathbf{x})$. The idea is that an infinitesimal variation $\tilde{f}(\mathbf{x})$ of the function $f(\mathbf{x})$ leads to a change in the objective function that is a linear functional of the perturbation $\tilde{f}$:

$$\Psi[f+\tilde{f}] \approx \Psi[f] + \int \frac{\delta\Psi}{\delta f}(\mathbf{x})\, \tilde{f}(\mathbf{x}) \mathrm{d}^N x\,. \tag{8.2}$$

This approximation is structurally identical to a Taylor expansion of the functional Ψ to first order. In this interpretation, the *variational derivative* $\frac{\delta\Psi}{\delta f}(\mathbf{x})$ plays the role of a gradient.

If we want to choose a perturbation $\tilde{f}$ of fixed small norm $N[\tilde{f}] = \varepsilon$ such that the change of the objective Ψ is maximal, we are essentially facing a constrained optimization problem for $\tilde{f}$. The problem can thus be tackled using Lagrange multipliers, according to which we need to find stationary points of the Lagrange function

$$\mathcal{L} = \Psi[f+\tilde{f}] - \Psi[f] - \lambda N[\tilde{f}] = \int \frac{\delta\Psi}{\delta f}(\mathbf{x})\tilde{f}(\mathbf{x})\mathrm{d}^N x - \lambda N[\tilde{f}]\,. \tag{8.3}$$

Here, λ is a Lagrange multiplier.

Assuming a quadratic form $N[\tilde{f}] = \frac{1}{2}\int \rho(\mathbf{x})\tilde{f}(\mathbf{x})^2 \mathrm{d}^N x$ of the norm (with $\rho(\mathbf{x}) > 0$), it is straight-forward to calculate the variational derivative of the Lagrange function $\mathcal{L}$ with respect to $\tilde{f}$ and equate it to zero:

$$\frac{\delta\mathcal{L}}{\delta\tilde{f}} = \frac{\delta\Psi}{\delta f}(\mathbf{x}) - \lambda\rho(\mathbf{x})\tilde{f}(\mathbf{x}) = 0 \tag{8.4}$$

$$\iff \quad \tilde{f}(\mathbf{x}) = \frac{1}{\lambda\rho(\mathbf{x})}\frac{\delta\Psi}{\delta f}(\mathbf{x})\,. \tag{8.5}$$

From the structure of the linear approximation (8.2) of Ψ it is clear that if we choose $\tilde{f}$ according to (8.5) with a positive value for λ, the objective function will

become larger. Thus, gradient descent on the objective function Ψ is done by choosing $\partial_\tau f(\mathbf{x}) = -\frac{\eta}{\rho(\mathbf{x})}\frac{\delta\Psi}{\delta f}(\mathbf{x})$, where now, $\eta := 1/\lambda > 0$ plays the role of a learning rate. Note that $\rho(\mathbf{x})$ enters the learning dynamics, so that different norms N for the variation $\tilde{f}$ lead to different learning rules. Because ρ is positive, however, the stationary solutions of the system, i.e., the fixed points of the objective Φ remain the same. Mathematically, the function $\rho(\mathbf{x})$ plays the role of an integrating factor (see, e.g., Reif & Muschik, 1987).

We will use the constraint that the variance of the variation δf is fixed: $\rho(\mathbf{x}) = p_{\mathbf{x}}(\mathbf{x})$. The resulting gradient descent learning rule is then given by

$$\partial_\tau f(\mathbf{x}) = -\frac{\eta}{p_{\mathbf{x}}}\frac{\delta\Psi}{\delta f}(\mathbf{x})\,. \tag{8.6}$$

Slowness Generates Diffusion

The optimization problem of SFA is formulated with an asymmetric decorrelation. From the perspective of biological plausibility, this is questionable, because there is no reason, why neurons of the same class in the same cortical area should undergo different adaptation mechanisms. Although neurons are never identical and there will always be a certain variability in their lateral connectivity pattern, it is unlikely that this variability implements the very specific form of asymmetry required by SFA. We believe that it may be more plausible to formulate the objective function Φ for a learning paradigm in a way that it is symmetrical, i.e., invariant with respect to the interchanges of the functions g_j. It should be investigated, however, to what extent the solutions of the system are robust to random disruptions of this symmetry before conclusions are drawn from this approach. In the following, we will gradually modify the objective functions to incorporate first the slowness objective and later also the constraints.

For slowness, a symmetrical objective can be formulated by simply summing over the Δ-values of all functions g_j:

$$\Phi[\mathbf{g}] = \sum_j \Delta(g_j)\,. \tag{8.7}$$

As already shown in chapter 3, the Δ-value of a function g_j can be written as a quadratic functional of the gradient $\partial_\mu g_j$. Using equation (3.12), the slowness objective (8.7) takes the form

$$\Phi[\mathbf{g}] \overset{(3.4,3.12)}{=} \sum_j \sum_{\mu,\nu} \int p_{\mathbf{x}}(\mathbf{x}) K_{\mu\nu}(\mathbf{x}) [\partial_\mu g_j(\mathbf{x})][\partial_\nu g_j(\mathbf{x})]\, \mathrm{d}^N x\,, \tag{8.8}$$

where $p_{\mathbf{x}}$ is the probability density of the input signals and $K_{\mu\nu}$ is the matrix of the second moments of the input velocity, conditioned on the input signal $\mathbf{x}$ (for definitions see equations (3.1) and (3.11)).

Because the objective Φ is a sum of Δ-values, each of which depends on one of the functions g_j only, its variational derivative with respect to a particular function g_j is given by the variational derivative of its Δ-value $\Delta(g_j)$. This derivative has already been calculated for the theory presented in chapter 3 (see appendix A). The resulting

dynamical equation for the gradient descent is given by[1]:

$$\partial_\tau g_j = \sum_{\mu,\nu} \frac{\eta}{p_{\mathbf{x}}} \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j \tag{8.9}$$

$$= \eta \sum_{\mu,\nu} \partial_\mu K_{\mu\nu} \partial_\nu g_j + \eta \sum_{\mu,\nu} [\partial_\mu \ln p_{\mathbf{x}}] K_{\mu\nu} \partial_\nu g_j \tag{8.10}$$

$$\overset{(3.18)}{=} -\eta \, \mathcal{D} g_j \, . \tag{8.11}$$

The learning dynamics (8.10) have the form of a partial differential equation and thus require boundary conditions for a unique solution. For SFA, the appropriate boundary condition (3.24) was shown to be of von Neumann type. It is reasonable to choose the same boundary condition for the gradient descent approach (8.10):

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j = 0 \quad \forall \mathbf{x} \in \partial V \, . \tag{8.12}$$

Structurally, equation (8.10) is a drift-diffusion equation with an inhomogeneous diffusion tensor $\eta \, K_{\mu\nu}(\mathbf{x})$. Thus, the dynamics of gradient descent tends to "smoothen" the functions g_j. Localized functions will become less localized with time. Because the diffusion tensor is essentially the matrix $K_{\mu\nu}$ and thus reflects the second-order moments of the input velocity, diffusion is faster in directions, where the input data vary quickly. From the perspective of slowness, this is perfectly reasonable, because smoothly varying functions will generate slower signals. The drift vector field in equation (8.10) is related to the gradient of the local entropy density $\ln p_{\mathbf{x}}$ of the distribution $p_{\mathbf{x}}$. This requires further examination, because it might provide a relation to information-theoretic methods. In the language of diffusion systems, the boundary condition (8.12) corresponds to reflecting boundaries.

The problem with this approach is that the diffusion will ultimately lead to homogeneous solutions, i.e., all functions g_j will be constant. With diffusion or slowness alone, the neurons thus fail to develop a tuning to stimulus features, instead they tend to lose all their selectivity. This is not surprising, because in SFA, the trivial constant solution had to be artificially excluded. To avoid the constant solution, constraints such as unit variance or decorrelation need to be added. How this can be done is discussed in section 8.3. First, we will show that an equation of the same class as equation (8.10) can also be obtained from the learning dynamics of STDP.

8.2.2 STDP: A Drift-Diffusion Approach

STDP in a Linear Poisson Neuron

We will use the same neuron model as in chapter 7: A linear Poisson neuron receives input spike trains $S_i(t)$ that are inhomogeneous Poisson processes with time-dependent firing rates $r_i(t)$. The output spike train $S(t)$ is also an inhomogeneous Poisson process with a firing rate that is given by the "membrane potential" $m(t) = \sum_j w_j [\epsilon \circ S_i](t)$.

[1] Here, we ignore the boundary integral that occurs in the derivation of the variational derivative and leads to the boundary condition (3.24) for the optimal functions of SFA (see Appendix A). This boundary term would introduce additional δ-shaped terms in the learning dynamics and implement von Neumann boundary conditions for the solutions. Instead, we require the boundary condition "by hand", so that the boundary integral vanishes.

Here, $\epsilon(t)$ reflects the shape of the EPSP and $\epsilon \circ S_i$ is the convolution of the EPSP with the input spike trains. Without loss of generality, we assume that $\int \epsilon(t)\mathrm{d}t = 1$.

Under the ergodicity assumption that temporal averaging can be replaced by ensemble averaging, the learning dynamics for STDP can be written as a linear dynamical system for the weights (Gerstner & Kistler, 2002, see also chapter 7):

$$\partial_\tau w_i = \sum_j \left[\int W_0(t) \langle r_j(t') r_i(t'+t) \rangle_{t'} \mathrm{d}t \right] w_j \,, \tag{8.13}$$

where $W_0 = \epsilon \circ W$ again denotes the effective learning window, i.e., the convolution of the EPSP and the learning window of STDP (see section 7.3.2, equation (7.54)). For simplicity, we have neglected spike-spike correlations, which would lead to an additional term (the second term in equation (7.52)). Spike-spike correlations can be neglected under the assumption that the number of input signals is very large, so that a single input channel contributes only weakly to the total output firing rate (Gerstner & Kistler, 2002).

Receptive Fields

The goal of this section is to map the weight dynamics (8.13) to a dynamical equation for the firing rate of the neuron as a function of a stimulus, i.e., for its receptive field. At this point, it should be noted that we use a generalized concept of a neuron's "receptive field". In constrast to the initial definition of the receptive field as the region in the sensory periphery in which a stimulation leads to a neuronal response (Sherrington, 1906), we incorporate the dependence of the firing rate of the neuron on the structure of the stimulus into the concept of the receptive field. In those applications that we have in mind, the stimulus is parameterized by a few (reduced) stimulus dimensions. In this approach, the receptive field of a visual neuron could be the dependence of its firing rate on the orientation and the spatial frequency of a visual grating, for example. Later in the chapter, we will also mention of the *size* of the RF. In our scheme, this means the extension of the region in the stimulus space, in which the firing rate of the neuron differs significantly from its spontaneous firing rate.

Thus, we assume that the stimulus can be characterized by a finite-dimensional vector $\mathbf{x}$. To stay close to the notation of the last section, let us denote the firing rate of the output neuron as a function of the stimulus with $g(\mathbf{x})$. The firing rate $g(\mathbf{x})$ is determined by the firing rates of the input neurons in response to the stimulus and the synaptic weights. Assuming that we know the receptive fields $R_i(\mathbf{x})$ of the input neurons and that the stimulus varies slowly compared to the time scale of the EPSP, the output firing rate is given by

$$g(\mathbf{x}) = \sum_i w_i R_i(\mathbf{x}) \,, \tag{8.14}$$

From this, we can immediately derive an equation for the dynamics of the receptive field of the output neuron:

$$\partial_\tau g(\mathbf{x}) = \sum_i \left[\partial_\tau w_i\right] R_i(\mathbf{x}) \,. \tag{8.15}$$

For simplicity, we will assume that the input neurons respond instantaneously to the stimulus, so that the input firing rate as a function of time is given by $r_i(t) = R_i(\mathbf{x}(t))$. Then, the correlation function between the input firing rates can be written in terms of

a two-point probability distribution $p_t(\mathbf{x}', \mathbf{x}''; t)$ of the stimuli and the receptive fields of the input neurons:

$$\langle r_j(t') r_i(t'+t) \rangle_{t'} = \iint \underbrace{\langle \delta(\mathbf{x}' - \mathbf{x}(t')) \delta(\mathbf{x}'' - \mathbf{x}(t'+t)) \rangle_{t'}}_{=: p_t(\mathbf{x}', \mathbf{x}''; t)} R_i(\mathbf{x}'') R_j(\mathbf{x}') \mathrm{d}^N x' \, \mathrm{d}^N x'' \,. \quad (8.16)$$

$p_t(\mathbf{x}', \mathbf{x}''; t)$ is the joint probability that the stimulus takes the values $\mathbf{x}'$ and $\mathbf{x}''$, with a temporal difference of t. Let us furthermore introduce a function $B(\mathbf{x}, \mathbf{x}')$, which is the temporal integral of the stimulus distribution, weighted with the effective learning window W_0:

$$B(\mathbf{x}', \mathbf{x}'') = \int W_0(t) p_t(\mathbf{x}', \mathbf{x}''; t) \, \mathrm{d}t \,, \quad (8.17)$$

which captures the statistical dependence of $\mathbf{x}'$ and $\mathbf{x}''$ on the time scale of the learning window, i.e., of tens of milliseconds. The reduced stimulus distribution $B(\mathbf{x}', \mathbf{x}'')$ will have significant contributions only when $\mathbf{x}' - \mathbf{x}''$ is so small that the stimulus can change from $\mathbf{x}'$ to $\mathbf{x}''$ within the time scale of the learning window, i.e., within tens of milliseconds. Thus, $B(\mathbf{x}', \mathbf{x}'')$ will generally be strongly localized in $\mathbf{x}' - \mathbf{x}''$.

Receptive Field Dynamics

Using the above expressions, we can write the learning dynamics (8.15) in terms of the receptive fields R_i:

$$\partial_\tau g(\mathbf{x}) \overset{(8.15)}{=} \sum_i (\partial_\tau w_i) R_i(\mathbf{x}) \quad (8.18)$$

$$\overset{(8.13,8.16,8.17)}{=} \sum_i \Big(\iint B(\mathbf{x}', \mathbf{x}'') R_i(\mathbf{x}'') \underbrace{\Big[\sum_j R_j(\mathbf{x}') w_j \Big]}_{\overset{(8.14)}{=} g(\mathbf{x}')} \mathrm{d}^N x' \, \mathrm{d}^N x'' \Big) R_i(\mathbf{x})$$

$$= \iint B(\mathbf{x}', \mathbf{x}'') \underbrace{\Big[\sum_i R_i(\mathbf{x}'') R_i(\mathbf{x}) \Big]}_{=: E(\mathbf{x}, \mathbf{x}'')} g(\mathbf{x}') \, \mathrm{d}^N x' \, \mathrm{d}^N x'' \quad (8.19)$$

$$\overset{(8.21)}{=} \int A(\mathbf{x}, \mathbf{x}') g(\mathbf{x}') \, \mathrm{d}^N x' \,, \quad (8.20)$$

where we defined

$$A(\mathbf{x}, \mathbf{x}') := \int B(\mathbf{x}', \mathbf{x}'') E(\mathbf{x}, \mathbf{x}'') \, \mathrm{d}^N x'' \,. \quad (8.21)$$

The learning dynamics (8.20) are a linear integro-differential equation. The central role in the dynamics is played by the integral kernel $A(\mathbf{x}, \mathbf{x}')$, which is in turn determined by two components: The reduced stimulus distribution $B(\mathbf{x}, \mathbf{x}'')$ and a term $E(\mathbf{x}, \mathbf{x}'')$ that measures "how often" the stimuli $\mathbf{x}$ and $\mathbf{x}''$ both evoke a response in the same presynaptic neuron and how strong these responses are. We already discussed that the former should be localized in $\mathbf{x}' - \mathbf{x}''$. The degree of localization of the latter in $\mathbf{x} - \mathbf{x}''$ reflects the size of the RF of the input neurons. If the input RF are localized, E will be localized in the difference of its arguments. If both p and E are localized in the difference of their arguments, A will also be localized in $\mathbf{x} - \mathbf{x}'$.

Two limit cases can be distinguished. Firstly, the RFs of the input neurons are so small that the stimulus leaves the receptive field on a time scale that is shorter than the

width of the learning window, i.e., within tens of milliseconds. In this case E would be more localized than p and the width of A would primarily be determined by the width of the reduced correlation function p. This case will occur mainly in stimulus directions that change extremely quickly and may moreover be problematic in the light of the assumption that the input neurons process their input instantaneously. For this reason, we expect that the width of the integral kernel A will in general be determined by E and thus by the size of the RFs of the input neurons.

Note also that E is positive semi-definite and symmetric in its arguments and that this is not necessarily the case for p.

Drift and Diffusion in Receptive Field Dynamics

We have discussed the localization of the integral kernel A for a particular reason: If A is localized in $\mathbf{x} - \mathbf{x}'$ on a spatial scale that is shorter than the typical scale of variation in $g(\mathbf{x})$, we can use a Kramers-Moyal expansion (see, e.g., Gardiner, 1985) and replace the integral equation (8.20) by a drift-diffusion equation. This is done by expanding $g(\mathbf{x}')$ in a second-order Taylor series around $\mathbf{x}$ and inserting this into equation (8.20):

$$\partial_\tau g(\mathbf{x}) \overset{(8.20)}{=} \int A(\mathbf{x},\mathbf{x}') g(\mathbf{x}')\,\mathrm{d}^N x' \tag{8.22}$$

$$\approx \int A(\mathbf{x},\mathbf{x}') \Big[g(\mathbf{x}) + \sum_\mu (x'_\mu - x_\mu)\partial_\mu g(\mathbf{x}) + \frac{1}{2}\sum_{\mu,\nu}(x'_\mu - x_\mu)(x'_\nu - x_\nu)\partial_\mu\partial_\nu g(\mathbf{x})\Big]\,\mathrm{d}^N x' \tag{8.23}$$

$$= M^{(0)}(\mathbf{x}) g(\mathbf{x}) + \sum_\mu M^{(1)}_\mu(\mathbf{x})\partial_\mu g(\mathbf{x}) + \sum_{\mu,\nu} M^{(2)}_{\mu\nu}(\mathbf{x})\partial_\mu\partial_\nu g(\mathbf{x})\,, \tag{8.24}$$

where we introduced the moments $M^{(i)}$ of the integral kernel A according to

$$M^{(0)}(\mathbf{x}) := \int A(\mathbf{x},\mathbf{x}')\,\mathrm{d}^N x' \overset{(8.21,8.17,8.19)}{=} \left(\int W_0(t)\mathrm{d}t\right) \sum_i \langle R_i\rangle_{\mathbf{x}} R_i(\mathbf{x})\,, \tag{8.25}$$

$$M^{(1)}_\mu(\mathbf{x}) := \int A(\mathbf{x},\mathbf{x}')(x'_\mu - x_\mu)\,\mathrm{d}^N x'\,, \tag{8.26}$$

$$M^{(2)}_{\mu\nu}(\mathbf{x}) := \frac{1}{2}\int A(\mathbf{x},\mathbf{x}')(x'_\mu - x_\mu)(x'_\nu - x_\nu)\,\mathrm{d}^N x'\,. \tag{8.27}$$

The RF dynamics (8.24) consist of three terms. The first term generates an exponential growth/decay of the RF g with a growth rate that is determined by the spatial structure of the 0-th moment of the kernel A. Note that if the input neurons are homogeneously distributed in the sense that they all fire with the same average firing rate and that their RF cover the input space homogeneously, the 0-th moment $M^{(0)}$ is independent of $\mathbf{x}$ (cf. equation (8.25)). In this case, all responses g grow/decay with the same rate, so this term would, by itself, not lead to structure formation of the RF. Note however, that it may well lead to structure formation by interacting with weight-limiting mechanisms. If $M^{(0)}$ is not homogeneous, it introduces a competition between the stimuli. Responses to stimuli with larger values for $M^{(0)}$ will grow faster. The exponential growth induced by the first term indicates that the dynamics (8.24) are usually unstable: they either diverge or lead to a extinction of the RF g. This is not surprising, however, because Hebbian learning without additional weight-limiting mechanisms is intrinsically unstable.

The second term of the RF dynamics creates a systematic drift of the RF along the vector field $M_\mu^{(1)}$. Such a drift of RFs has been observed both in experiments and in simulations. For example, M. R. Mehta et al. (1997) observed that hippocampal place fields in linear tracks tend to shift, when the rat traverses the track repeatly in the same direction. Yao & Dan (2001) have shown that the orientation tuning of cells in primary visual cortex can be shifted by repeated presentation of uni-directionally rotating gratings. In addition, simulations have shown that systematic shifts in place field position can be explained by STDP type learning rules (Abbott & Blum, 1996; M. Mehta et al., 2000). In all these studies, the stimulus statistics were irreversible in the sense that the stimuli followed designated paths unidirectionally. The results of these studies indicate that RF drift occurs such that the response of the neuron, interpreted according to its response pattern before learning, codes for stimuli that are expected to occur at a later moment in time. Thus, this kind of RF plasticity can be interpreted in terms of a prediction of future events. Further research will be necessary to investigate if the drift term in the RF dynamics (8.24) can be interpreted in terms of predictive coding.

The third term in the RF dynamics introduces a diffusion of the RF according to an inhomogeneous diffusion tensor $M_{\mu\nu}^{(2)}$. If $M_{\mu\nu}^{(2)}$ is positive definite (cf. the discussion below), this term tends to make the RFs larger, particularly in directions of eigenvectors of $M_{\mu\nu}^{(2)}$ with large eigenvalues. In these directions, the dynamics tend to flatten the response pattern of the cell. They can thus be interpreted as a means of invariance learning. Note that, in contrast to gradient-based slowness learning, the directions in which the RFs tend to become larger is not purely determined by the directions in which the stimulus changes quickly. The diffusion tensor $M_{\mu\nu}^{(2)}$ measures the delocalization of the integral kernel A, which will in general – as discussed above – be determined by the size of the RFs of the input neurons. An expansion of hippocampal place fields that is consistent with such a diffusion component of Hebbian plasticity has been shown both experimentally and in simulations (M. R. Mehta et al., 1997; M. Mehta et al., 2000).

Note that the integral kernel A is not necessarily positive definite, mainly because p may not be positive definite. Thus, the diffusion tensor $M_{\mu\nu}^{(2)}$ may have negative eigenvalues. Consequently, anti-diffusion, i.e., a contraction of the RF in directions of eigenvectors of of $M_{\mu\nu}^{(2)}$ with negative eigenvalues cannot be excluded. Whether this occurs is strongly influenced by the shape of the learning window. For example, a strictly negative learning window would lead to a diffusion tensor that is *negative* definite. This makes further analysis of the influence of the shape of the learning window on the diffusion tensor rather interesting, because it might provide indirect evidence for the shape of the physiologically relevant component of the learning window. For example, it has been argued that for reasons of stability, the LTD component should dominate over the LTP component (Song & Abbott, 2001). Also, physiological studies have provided evidence that the LTD branch of STDP may decay on a slower time scale than the LTD branch (Feldman, 2000). An examination whether these learning-window shapes lead to diffusion or contraction of the RF could lead to experimentally testable predictions for the dynamics of RF.

Our analysis also raises the possibility of measuring the learning window indirectly, by measuring the RF dynamics while artificially manipulating the input statistics. The main problem in this approach is that the receptive fields of the input neurons also enter the learning dynamics, and these may be hard to determine. Still, this approach may be promising for the study of synaptic plasticity in cortical areas in which receptive field properties are relatively well-characterized, e.g., in primary visual cortex. At least one

experiment that points in this direction has already been reported: Yao & Dan (2001) studied temporal requirements for the input statistics for the plasticity of orientation tuning in primary visual cortex and compared these requirements with the time course of the STDP learning window.

8.3 The Role of Constraints

8.3.1 Reaction-Diffusion Systems

A Reaction-Diffusion System with Global Coupling

We already discussed that the diffusion of the RF counteracts the development of stimulus specific responses, it tends to "flatten" the RF. This indicates that the learning dynamics may have a single stationary solution. In this case, if all neurons follow the same RF dynamics, they will ultimately encode the same stimulus features. This can be avoided by the introduction of constraints such as unit variance and decorrelation.

To incorporate constraints in a gradient descent, it is common to introduce additional *punishment terms* in the objective function (8.7) that lead to positive contributions when the functions $\mathbf{g}$ fail to fulfill the constraints. An appropriate choice of punishment terms Φ_U, Φ_D for unit variance and decorrelation is given by:

$$\Phi_U[\mathbf{g}] = \sum_j \sigma_U \left(\langle g_j^2 \rangle_{\mathbf{x}} - 1 \right) , \tag{8.28}$$

$$\Phi_D[\mathbf{g}] = \sum_{i,j \neq i} \sigma_V \left(\langle g_i g_j \rangle_{\mathbf{x}} \right) . \tag{8.29}$$

where $\sigma_{U/V}(z)$ are positive point nonlinearities that vanish for $z = 0$, and which are monotonicly increasing with $|z|$, e.g., $\sigma(z) = z^2$. Note that, due to the lack of a zero mean constraint, Φ_U does not implement a unit variance constraint. The punishment term Φ_U vanishes not only for functions with unit variance, but also for the constant $g_j = 1$. Therefore, the constant solution cannot be avoided by the "unit variance" punishment term Φ_U alone. In combination with the decorrelation term Φ_D, however, the homogeneous constant solutions can become unstable, as discussed below.

An objective function that enforces slowness while taking the constraints into account is given by

$$\Phi[\mathbf{g}] = \sum_j \Delta(g_j) + \frac{\alpha}{2} \Phi_U[\mathbf{g}] + \frac{\beta}{2} \Phi_D[\mathbf{g}] , \tag{8.30}$$

where α and β are trade-off parameters that weight the importance of the constraints.

The associated gradient descent learning rule is given by

$$\partial_\tau g_j \overset{(8.6,8.11)}{=} -\eta \mathcal{D} g_j - \eta \frac{\alpha}{2p_{\mathbf{x}}} \frac{\delta \Phi_U}{\delta g_j} - \eta \frac{\beta}{2p_{\mathbf{x}}} \frac{\delta \Phi_D}{\delta g_j} \tag{8.31}$$

$$= -\eta \mathcal{D} g_j - \underbrace{\eta\, \alpha\, \sigma_U' \left(\langle g_j^2 - 1 \rangle_{\mathbf{x}} \right)}_{=: F\left(\langle g_j^2 - 1 \rangle_{\mathbf{x}}\right)} g_j - \eta\, \beta \sum_{i \neq j} \underbrace{\sigma_V' \left(\langle g_i g_j \rangle_{\mathbf{x}} \right)}_{=: J_{ji}} g_i \,. \tag{8.32}$$

$$= -\eta \mathcal{D} g_j - F\left(\langle g_j^2 - 1 \rangle_{\mathbf{x}} \right) g_j - \sum_{i \neq j} J_{ji} g_i \,. \tag{8.33}$$

Structurally, equation (8.33) has the form of a reaction-diffusion equation with global coupling. Global coupling in this context means that the spatial averages $\langle g_j^2 \rangle_{\mathbf{x}}$ and $\langle g_i g_j \rangle_{\mathbf{x}}$ have an influence on the local dynamics of the functions g_j. Reaction-diffusion system are somewhat canonical in the theoretical description of spatiotemporal pattern formation. Thus, the concepts provided in this chapter may serve as a bridge from models of synaptic plasticity to the theory of pattern formation. This raises the interesting possibility of describing the properties of receptive fields within the well-developed theoretical framework of nonlinear dynamics.

A problem with the interpretation of equation (8.33) as a reaction-diffusion system is that in order to remain close to SFA, we have not restricted the functions g_j to be non-negative, which is typically the case in reaction-diffusion systems. Thus, they cannot be interpreted as chemical concentrations or populations. It is worth noting, though, that in the light of biological plausibility, it is highly reasonable to restrict the output signals g_j to be positive, because this allows an interpretation in terms of firing rates.

Speculations on Biological Mechanisms

The reaction-diffusion system (8.33) can be interpreted in terms of biologically feasible mechanisms. Let us think of the functions g_j as responses of cortical neurons.

The second term, which arises from the unit variance constraint tends to multiplicatively scale the response of the neurons, depending on the deviation of the mean square of the response from a target value, in this case the target value one. Qualitatively, this resembles a homeostatic mechanism that aims at keeping the neuron's average response within a target range. Homeostatic plasticity has been found in a number of experimental preparations, either as a multiplicative rescaling of synaptic efficacies (synaptic scaling, Turrigiano et al., 1998) or as a dynamic regulation of neuronal excitability (Desai et al., 1999). For a recent review on homeostatic plasticity see (Turrigiano, 2007). The requirement that the mean square of the output signal remains close to a target value, motivated by the unit variance constraint, appears arbitrary in the context of homeostasis. It may well be possible, however, that other objectives that aim at stabilizing the mean firing rate instead, lead to qualitatively similar results (see the qualitative discussion of the stationary solutions below). It would be interesting to investigate how sensitive the qualitative behavior of a system with slowness/diffusion and homeostasis is with respect to model details.

The third term in equation (8.33) arises from the decorrelation constraint. It mediates an "inhibitory" coupling between the response properties of the neurons with a

coupling strength J_{ij} that depends on the correlation of their output signals. The coupling strength J_{ij} could thus be learned by a Hebbian mechanism for the synaptic weights of recurrent connections. Hebbian learning on recurrent connections is rather common in the literature, not only for decorrelation purposes (Földiàk, 1989; Barlow & Földiák, 1989) but also in other contexts, e.g., for self-organizing feature maps (Bednar & Miikkulainen, 2000)[2]. The network implementation of principal subspace analysis proposed by Földiák (Földiàk, 1989) uses this paradigm: Excitatory input connections to a network of linear neurons are learned by Oja's rule (Oja, 1982) and inhibitory recurrent connections are adapted according to Hebbian learning. The inhibitory recurrence reduces the responses of neurons whose output signals are significantly correlated, which in turn reduces the adaptation of those neurons to the input patterns. In contrast, neurons whose output signals are uncorrelated to the other neurons are more active and can thus adapt more efficiently. Uncorrelated neurons thus have an "advantage" during learning, so the synaptic weights of such a network stabilize into a state where the output signals are uncorrelated.

In summary, all three components of the learning rule (8.33) should be implementable within the constraints of neuronal circuitry.

Qualitative Discussion of the Stationary Solutions

Depending on the trade-off parameters α and β, the stationary solutions to the reaction-diffusion system (8.33) and the optimal functions for SFA can differ more or less strongly.

For small β, the decorrelation constraint may not be sufficiently enforced. In this case, the stationary solutions of equation (8.33) are spatially homogeneous, because, both the slowness objective and the punishment term Φ_U for "unit variance" vanish for $g_j = 1$.

With increasing β, the value of the punishment term becomes larger, which makes the homogeneous steady state less and less favorable. At some critical value for β, decorrelation will become so strong that the homogeneous solution becomes unstable and spatially patterned responses emerge: The "neurons" develop stimulus selectivity. In the limit case of very large β, the stationary solutions will be almost perfectly uncorrelated. Small deviations from unit variance and decorrelation may remain, however, to trade off for a smaller Δ-value. Thus, for sufficiently strong enforcement of the constraints, the solutions to SFA (including the constant function) should be stationary solutions to the gradient descent system (8.33). Note however, that because decorrelation is symmetric in the gradient descent approach, any orthogonal mixture of the solutions to SFA is also a solution to equation (8.33).

8.3.2 Temporally Restricted Constraints

As discussed above, the reaction-diffusion system (8.33) can be interpreted in terms of biologically plausible mechanisms. There is one detail, however, that may be difficult to implement: global coupling. The problem can be illustrated by means of the homeostatic term: The learning rule introduces a "multiplicative scaling" of the functions g_j that depends on the spatial average of the square of g_j. The crux is that the average is done over *all* possible input values $\mathbf{x}$. The calculation of this average requires the neuron to

[2]There are indications, however, that Hebbian learning for intracortical connections is not purely correlation-based, but rather subject to an STDP type learning rule (Yao et al., 2004; Young et al., 2007).

track the variance of its output over a time scale that is long enough to sample every single point in the environment.

The time scale of homeostatic plasticity, although still under investigation, was mostly thought to be on the order of days (see, e.g., Desai et al., 1999; Murthy et al., 2001). However, recent results from the visual system of goldfish indicate that homeostasis may occur on a much shorter time scale on the order of an hour or less (Riegle & Meyer, 2007). No matter, what the exact time scale of homeostasis is, it will be finite, so that events in the distant past will be less important than recent events. From this perspective, the global spatial average $\langle\cdot\rangle_{\mathbf{x}}$ may no longer be appropriate. Rather, the averaging needed to calculate the variance of the output signal should be done over a timescale τ_U that is associated with that of homeostasis. One possibility of implementing this time scale is by tracking the variance of a neuron in terms of an exponential trace $v(t)$:

$$v_j(t) = \int_{\infty}^{t} \tau_U^{-1} \exp\left(-\frac{t-t'}{\tau_U}\right) g_j(\mathbf{x}(t'))^2 \mathrm{d}t' \,. \tag{8.34}$$

Such a trace can be implemented in a biologically plausible fashion by means of a leaky integrator. We can now easily define a new punishment term that aims at keeping the trace close to unity:

$$\Phi'_U[\mathbf{g}] = \sum_j \langle (v_j(t)-1)^2 \rangle_t \,. \tag{8.35}$$

To make this functional compatible with the formalism introduced before, we need to formulate it in terms of probability densities. To do so, we introduce a joint probability density $B(\mathbf{x}', t'; \mathbf{x}'', t'')$ that quantifies the probability that the input takes the value $\mathbf{x}'$ at time t' and the value $\mathbf{x}''$ at time t''. Under an ergodicity assumption, we can then rewrite the punishment term (8.35) by inserting an additional ensemble average, weighted with p:

$$\begin{aligned}
\Phi'_U[\mathbf{g}] &= \sum_j \int_V \int_V \mathrm{d}^N x' \, \mathrm{d}^N x'' \left(g_j(\mathbf{x}')^2 - 1\right)\left(g_j(\mathbf{x}'')^2 - 1\right) \times \\
&\quad \times \underbrace{\left\langle \int_{-\infty}^{t} \int_{-\infty}^{t} \tau_U^{-2} B(\mathbf{x}', t'; \mathbf{x}'', t'') \exp\left(-\frac{2t - (t'+t'')}{\tau_U}\right) \mathrm{d}t' \mathrm{d}t'' \right\rangle_t}_{=:\varphi_U(\mathbf{x}', \mathbf{x}'')} \qquad (8.36) \\
&= \sum_j \int_V \int_V \varphi_U(\mathbf{x}', \mathbf{x}'') \left(g_j(\mathbf{x}')^2 - 1\right)\left(g_j(\mathbf{x}'')^2 - 1\right) \mathrm{d}^N x' \, \mathrm{d}^N x'' \,. \qquad (8.37)
\end{aligned}$$

The structure of the interaction function φ_U is determined by two factors: the time scale τ_U and the distribution p. First note that only times t' and t'' significantly contribute to φ_U, for which $2t - (t' + t'')$ is smaller or on the order of τ_U. Because $t' < t$ and $t'' < t$, this is only possible if both times are relatively close to t. Thus, $t' - t''$ should be on the order of τ_U, as well. Secondly, for small time differences $t' - t''$, the probability distribution p should be localized in $\mathbf{x}' - \mathbf{x}''$, because the (continuous) input signal can only change by a limited amount within a given time $t' - t''$. In summary: If τ_U is small, only small values of $t' - t''$ are important for the interaction function φ_U. However, for small values of $t' - t''$, the joint density $B(\mathbf{x}', \mathbf{x}'')$ is localized in $\mathbf{x}' - \mathbf{x}''$. Thus, $\varphi_U(\mathbf{x}', \mathbf{x}'')$ should be localized in $\mathbf{x}' - \mathbf{x}''$ if τ_U is small. The degree of localization of φ_U is determined by the typical "distance" the input signal can "traverse" on the time scale of τ_U, or, to put it differently, by the ratio of the autocorrelation time of the input signals and the

timescale τ_U. Note that if τ_U is much larger than the autocorrelation time, the global constraints of the last section are recovered.

Similar considerations can be done for the decorrelation constraint, leading to a "localized" punishment term

$$\Phi'_D[\mathbf{g}] = \sum_{i,j(\neq i)} \int_V \int_V \varphi_V(\mathbf{x}', \mathbf{x}'') g_i(\mathbf{x}') g_j(\mathbf{x}') g_i(\mathbf{x}'') g_j(\mathbf{x}'') \mathrm{d}^N x \, \mathrm{d}^N x' \,. \tag{8.38}$$

Replacement of the punishment terms in equation (8.31) by their localized versions (8.37) and (8.38) yields a new learning rule:

$$\begin{aligned} \partial_\tau g_j(\mathbf{x}) = -\mathcal{D} g_j(\mathbf{x}) - \frac{\eta\,\alpha}{p_\mathbf{x}} \left[\int_V \varphi_U(\mathbf{x}, \mathbf{x}') \left(g_j(\mathbf{x}')^2 - 1 \right) \mathrm{d}^N x' \right] g_j(\mathbf{x}) \\ - \frac{\eta\,\beta}{p_\mathbf{x}} \sum_{i(\neq j)} \left[\int_V \varphi_V(\mathbf{x}, \mathbf{x}') g_j(\mathbf{x}') g_i(\mathbf{x}') \mathrm{d}^N x' \right] g_i(\mathbf{x}) \,. \end{aligned} \tag{8.39}$$

In the physics of pattern formation, spatial interactions are often classified as local, nonlocal or global, depending on their range. Local interactions mean that the dynamical equation for a field $g(\mathbf{x})$ contain only local quantities at "position" $\mathbf{x}$. In this case, the formation of spatial patterns is mediated by a "maximally localized coupling", mathematically reflected by differential operators that describe, e.g., diffusion. Nonlocal coupling means that quantities at other positions $\mathbf{x}'$ play a role in the local dynamics of $g(\mathbf{x})$, but that this influence declines with the distance $|\mathbf{x} - \mathbf{x}'|$. A typical example is the electromagnetic field whose strength declines with the inverse distance. Global coupling, finally, means that the range of the interaction is so much larger than the system size that practically all spatial positions have the same influence on the local dynamics. According to this classification, the learning rule (8.33) implements global coupling, because of the global spatial average. In contrast, the localized version (8.39) would be classified as nonlocal coupling.

What is the advantage of nonlocal coupling over global coupling, apart from being biologically more plausible? We believe that nonlocal constraints will reduce the dependence of the solutions on the boundary conditions. This dependence was disturbing, e.g., when we tried to interpret the grid-like structure of the solutions to SFA for the self-localization problem in section 4.2. The arrangement of high activity regions in the SFA simulations depended strongly on the boundary conditions and thus on the shape of the room. In a rectangular room, the solutions show rectangular grids, whereas, e.g., for circular rooms, the grids show a circular arrangement. This behavior is conflict with experimental results for grid cells in entorhinal cortex, which show a hexagonal arrangement of firing fields even in rectangular rooms (Hafting et al., 2005). For the model with local constraints (8.39) we expect that boundary effects will be negligible at distances from the boundaries that are larger than the spatial scale introduced by the interaction functions φ_U and φ_V, so that the final pattern in the middle of the room becomes independent of room shape. Another interesting question is, if the localization of the interaction functions introduces an intrinsic spatial scale in the receptive fields, which is missing for the solutions of SFA. The solutions of SFA show oscillations whose spatial frequency is determined by the shape of the stimulus space and the number of solutions that are taken into account. In contrast, grid cells (and also many other receptive fields

in the brain) have intrinsic spatial scales, since their grids show spatial frequencies that roughly cover the octave between 40 and 80cm (Hafting et al., 2005).

Hexagonal grids are a common stationary state in reaction-diffusion systems, suggesting that the developed framework may be an interesting starting point for a grid cell model.

8.4 Discussion

Online learning rules are biologically more plausible than batch learning. In addition, they have the advantage that they allow to track the dynamics of the learning process. Since the weights implicitly determine the receptive field of the neuron, it is then also possible to track the dynamics of the receptive field during learning. In this chapter, we have studied two systems and shown that their receptive field dynamics can be described in terms of drift-diffusion equations: Gradient-based slowness learning and STDP in linear Poisson neurons.

Hebbian learning is generally unstable: Either the weights diverge or they undergo extinction. Moreover, the diffusion term in the RF dynamics counteracts the development of stimulus selectivity. Without interactions between the output neurons, it is likely that they all develop the same receptive fields (unless the learning rule has several stable fixed points), which are moreover unselective and thus uninformative about the stimulus. To avoid these problems, we have studied the effect of constraints on the learning dynamics. Using a constrained gradient descent, we have shown that the unit variance constraint – which implements a stabilization the learning dynamics – can be interpreted in terms of homeostatic plasticity and that the decorrelation constraint can be implemented by means of inhibitory lateral interactions whose strength is subject to Hebbian learning. Note that in our approach, the development of stimulus selectivity is not built into the learning rule for the isolated neuron already (as done, e.g., by Bienenstock et al., 1982). Instead, the selectivity is caused by lateral interactions between the output neurons that force the neurons to code for different aspects of the stimulus.

The terms that arise from the constraints in the gradient-based approach could of course easily be transferred to the drift-diffusion system derived from STDP. From a paradigmatic point of view, this would be questionable, however, because the drift-diffusion equation for STDP was derived from a physiologically motivated model. Conceptually, it would thus be favorable to derive "constraint" terms directly from models of biological mechanisms such as homeostatic plasticity and lateral interactions. Building a model for homeostatic plasticity is probably simpler than for lateral interactions, although the associated delayed negative feedback loop may require careful modeling to avoid oscillations (Rossum et al., 2000). Biologically plausible modeling of the plasticity of lateral inhibition is a more complicated task, mainly because this interaction is mediated by interneurons. Interneurons in neocortex are rather diverse in their morphology and probably also in their function. Moreover, plasticity of inhibitory synapses or synapses onto inhibitory neurons is far from being fully understood. For example, STDP of synapses made by pyramidal cells onto inhibitory interneurons seems to depend on the target cell type (Lu et al., 2007), so there may not be a simple representative model for inhibitory synaptic plasticity. Building a biologically plausible but still simple implementation of this type of interaction will require careful study of the relevant physiological literature.

Clearly, the next step is to test if the dynamical equations we derived can account for the response behavior of cortical cells. For the slowness model, the simulation of the RF dynamics should be straight-forward: We have to choose a stimulus paradigm, e.g., orientation tuning in visual neurons, and model its input statistics. Then, the resulting stationary solutions of the receptive field dynamics can be studied in dependence of (a) the trade-off parameters α and β for the unit variance and decorrelation constraints, (b) the implementation of the constraints – global vs. nonlocal and (c) (possibly dynamic changes of) the input statistics.

For the STDP-based model, the implementation is not so simple. As discussed above, we first need to establish a biologically valid description of the constraints. The next problem is that not only the stimulus statistics have to be known but also the receptive fields of the input neurons. This model can thus only be tested for stimulus dimensions and neuron types for which the receptive fields of the upstream neurons are well studied. Candidate systems could be found, e.g., in the early visual system. Orientation tuning curves in V1 have been well-characterized. What kind of orientation preferences do we expect downstream? Projections from simple cells to complex cells could be modeled in a two-dimensional stimulus space of orientation and spatial phase of a grating. Can the learning dynamics account for the phase invariance and orientation selectivity of complex cells? What is the influence of the stimulus statistics on the response behavior of the cells? A drastic example: Can we design input statistics that turn complex cells, which are invariant with respect to spatial phase while being selective for orientation, into cells that are invariant to orientation but are selective for phase?

Several recent studies suggest that cortical receptive field plasticity is mediated not by plasticity of synapses originating from earlier processing stages, e.g., at thalamocortical synapses, but rather by intracortical plasticity (Yao et al., 2004; Froemke et al., 2007). In the light of these studies, receptive field dynamics should be studied in a framework of recurrent networks with plastic recurrent connections and static connections to external inputs. Conceptually, this problem is more complicated than "classical" feed-forward models, because changes in the recurrent synaptic weights change not only the receptive field of an output neuron, but also those of the input neurons, simply because all neurons play both roles. A change in the receptive fields of the input neurons, however, affects the parameters of the learning dynamics, so that the learning dynamics themselves become dynamical. As a consequence, the resulting equations will become highly nonlinear. In the stationary state, the receptive fields of all neurons in the network have to be mutually consistent. A change in the receptive field of a single neuron would change the dynamics of the others and thus lead to a different stationary state of the whole network.

An aspect that was neglected in our approach is the temporal structure of receptive fields. It would be interesting to examine if the full spatiotemporal receptive field of the input neurons can be incorporated into the drift-diffusion model of STDP and if the resulting receptive field dynamics reflect the suggested function of STDP in minimizing the response latency (Guyonneau et al., 2005).

In conclusion, this chapter has provided a formal bridge from gradient-based slowness learning and STDP to the theory of self-organized pattern formation. This link should allow to transfer interesting techniques and insights from nonlinear dynamics to studies of receptive field formation. First, however, the approach should be tested on exemplary systems to check if it can indeed describe receptive fields as found in the brain.

Chapter 9

Conclusion

Summary

The central objective of this thesis was to study the unsupervised learning principle of slowness from different perspectives. Two approaches were given particular emphasis: The mathematical analysis of slow feature analysis and the question, if slowness can be implemented through biologically plausible mechanisms.

In Part I we showed that SFA allows an in-depth analytical treatment that reveals analogies to well-studied physical systems and allows to make analytical predictions for concrete applications. In particular, we presented (a) a new algorithm for nonlinear blind source separation, (b) analytical predictions for SFA as part of a model for place and head-direction cells and (c) analytical results for SFA as a model for the self-organized formation of complex cell receptive fields. In chapter 6, we provided a link between slowness learning and predictive coding and discussed its limitations.

In the second part of the thesis, we examined the slowness principle with respect to a possible biological implementation. Having shown in chapter 7 that temporally nonlocal Hebbian plasticity can under certain conditions be interpreted as an gradient-based implementation of slowness learning, chapter 8 was dedicated to an analysis of receptive field dynamics under gradient-based slowness learning and spike-timing-dependent plasticity.

Direct Experimental Evidence for Slowness?

The observation that slowness can serve as a basis for models of neural response properties in so different systems like primary visual cortex (Einhäuser et al., 2002; Berkes & Wiskott, 2005) and the hippocampus (Wyss et al., 2006; Franzius, Sprekeler & Wiskott, 2007) together with the conceptual proof that it is implementable by biologically plausible means (Sprekeler et al., 2007) suggests that slowness may be a computational principle that is used in the brain. Unfortunately, most currently available studies on slowness learning can only be considered as indirect evidence for this conjecture, because neural response properties that can be reproduced with slowness learning may also be explainable by other computational principles. Therefore, it would be favorable to develop experimental paradigms that test explicitly if the slowness principle is at work in sensory systems. Such an experimental paradigm would require systematic manipulations of the temporal stimulus statistics and subsequent examination if neural response properties are differentially altered.

A few experiments that follow this line of thought have been performed for higher processing stages in the visual system. For example, psychophysical studies have shown that human subjects have a higher probability of confusing objects that were repeatedly viewed in quick temporal succession (Wallis & Bülthoff, 2001; Cox et al., 2005). Physiological evidence has been provided by Miyashita (1988), who showed that when monkeys were trained with a series of geometrical shapes in fixed order, responses of neurons in inferotemporal cortex became correlated among stimuli that occurred in temporal vicinity.

It would be interesting to study if neural correlates of the slowness principle can also be found in earlier stages of sensory processing. A possible experimental system for this question could be the orientation tuning of cells in primary visual cortex. According to the slowness principle, the sharpness of the orientation tuning should decrease with the speed of the rotation in the stimulus. This could be tested by confronting an animal with rotating gratings and correlating the mean (change in) sharpness of the orientation tuning curves with the mean squared angular velocity of the rotation.

This experimental paradigm is not restricted to visual orientation tuning. The same type of experiment could be done for any cell type that shows a well-characterized tuning to a stimulus dimension whose statistics can be controlled. One could for example study if the selectivity to spatial phase of simple cells in primary visual cortex can be weakened by quickly translating stimuli or if the frequency tuning of cells in the auditory system can be altered by artificial stimuli that contain frequency sweeps with different velocities.

In closing, I hope to have shown that slowness is an interesting computational principle, both for models of sensory processing and for applications in signal processing. I hope that this thesis has contributed to an understanding of slowness learning and that it will be of use for further applications.

Acknowledgements

There are many people without whom this work would not have been possible. First and foremost I would like to thank Laurenz Wiskott for his guidance and his support. Apart from giving me the chance of pursueing my research in excellent working conditions, he was always open for constructive, concentrated and enjoyable scientific discussions. I would like to thank him for inspiration, for constructive criticism and for giving me the freedom to pursue my own ideas. I learned a lot.

It was a great pleasure to work with all members of his group, especially with Mathias Franzius, with whom I share not only place cells, but also a liking for the good things in life. I am furthermore indebted to Tiziano Zito, for the enjoyable cooperation, for his extraordinary choice of music pieces about differential equations and for the uncountable times he cured my computer from the damage I had inflicted. At the keyboard, he is a magician. I would also like to thank our monocycling bin-scorer Niko Wilbert and Susanne Lezius for patiently reading and commenting the manuscript. Thanks also to Pietro Berkes for constructive discussions and for the permission to use his figures. It was also a great pleasure to work, discuss and celebrate with Felix Creutzig.

I had a great time at the ITB and I am grateful to all its members for contributing to the fantastic atmosphere at the institute. In particular I would like to thank Roland Schaette, Samuel Glauser, Jan Benda, Susanne Schreiber, Martin Stemmler, and Richard Kempter for music, coffee, discussions and teaching opportunities.

Outside of the ITB, I would like to thank Mark van Rossum for the great opportunity of spending a rainy, but enjoyable and interesting month at the ANC in Edinburgh. I also thank the reviewers of the articles for their constructive feedback.

The Volkswagen foundation funded large parts of this work.

Last, but not least, I would like to thank all my friends, my sister Nele, my parents and Nora.

Bibliography

Abbott, L. F. & Blum, K. I. (1996). Functional significance of long-term potentiation for sequence learning and prediction. *Cerebral Cortex*, *6*(3), 406-416.

Abbott, L. F. & Nelson, S. B. (2000). Synaptic plasticity: Taming the beast. *Nature Neuroscience*, *3*, 1178–1183.

Adelsen, E. H. & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal Optical Society of America A*, *2*(2), 284–299.

Albus, K. & Wolf, W. (1984). Early post-natal development of neuronal function in the kitten's visual cortex: a laminar analysis. *Journal of Physiology*, *348*, 153–185.

Barlow, H. & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In (S. 54–72). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Barndorff-Nielsen, O. & Cox, D. (1989). *Asymptotic techniques for use in statistics*. New York: Chapman and Hall.

Becker, S. & Hinton, G. E. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*(6356), 161–163.

Bednar, J. & Miikkulainen, R. (2000). Tilt aftereffects in a self-organizing model of the primary visual cortex. *Neural Computation*, *12*(7), 1721–1740.

Bell, A. J. & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bell, A. J. & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F. & Moulines, E. (1997). A blind source separation technique using second order statistics. *IEEE Transactions on Signal Processing*, *45*, 434–444.

Berkes, P. & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cells. *Journal of Vision*, *5(6)*, 579-602.

Berkes, P. & Zito, T. (2007). *Modular modular toolkit for data processing (version 2.1)*. http://mdp-toolkit.sourceforge.net.

Bi, G. & Poo, M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, *18*(24), 10464–10472.

Bienenstock, E. L., Cooper, L. N. & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*(1), 32–48.

Blaschke, T., Berkes, P. & Wiskott, L. (2006). What is the relation between slow feature analysis and independent component analysis? *Neural Computation*, *18*(10), 2495–2508.

Blaschke, T. & Wiskott, L. (2004, May). CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*, *52*(5), 1250–1256.

Blaschke, T., Zito, T. & Wiskott, L. (2007). Independent slow feature analysis and nonlinear blind source separation. *Neural Computation*, *19*(4), 994-1021.

Bray, A. & Martinez, D. (2002). Kernel-based extraction of slow features: Complex cells learn disparity and translation invariance from natural images. In *Advances in neural information processing systems 15* (S. 253–260). New York: MIT Press.

Brenner, N., Bialek, W. & Steveninck, R. de Ruyter van. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, *26*, 295–701.

Buonomano, D. V. & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Reviews of Neuroscience*, *21*, 149–186.

Cang, J., Renteria, R., Kaneko, M., Liu, X., Copenhagen, D. & Stryker, M. (2005). Development of precise maps in visual cortex requires patterned spontaneous activity in the retina. *Neuron*, *48*(5), 797–809.

Chechik, G., Globerson, A., Tishby, N. & Weiss, Y. (2005). Information bottleneck for Gaussian variables. *The Journal of Machine Learning Research*, *6*, 165-188.

Courant, R. & Hilbert, D. (1989). *Methods of mathematical physics Part I.* New York: Wiley.

Cox, D., Meier, P., Oertelt, N. & DiCarlo, J. (2005). "Breaking" position-invariant object recognition. *Nature Neuroscience*, *8*(9), 1145–1147.

Creutzig, F. (2008). *Sufficient encoding of dynamical systems.* Unveröffentlichte Dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät I, Universitätsbibliothek.

Creutzig, F. & Sprekeler, H. (2008). Predictive coding and the slowness principle: An information-theoretic approach. *Neural Computation*, *20*(4), 1026–1041.

Davydov, A. (1976). *Quantum mechanics.* New York: Pergamon Press.

Dayan, P., Häusser, M. & London, M. (2004). Plasticity kernels and temporal statistics. In *Advances in neural information processing systems 16.* New York: MIT Press.

De Valois, R. L., Yund, E. W. & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*, 531–544.

Debanne, D., Gähwiler, B. H. & Thomson, S. M. (1994). Asynchronous pre- and post-synaptic activity induces associative long-term depression in area CA1 of the rat hippocampus. *PNAS*, *91*, 1148–1152.

Desai, N., Rutherford, L. & Turrigiano, G. (1999). Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience*, *2*(6), 515–520.

Dimitrov, A. G. & Miller, J. P. (2001). Neural coding and decoding: Communication channels and decoding. *Network: Computation in Neural Systems*, *12*, 441-472.

Dong, D. W. (2001). Spatiotemporal inseparability of natural images and visual sensitivities. *Computational, neural & ecological constraints of visual motion processing*, 371.

Dong, D. W. & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, *6*(3), 345–358.

Drew, P. & Abbott, L. (2006). Extending the effects of spike-timing-dependent plasticity to behavioral timescales. *Proceedings of the National Academy of Sciences*, *103*(23), 8876–8881.

Einhäuser, W., Kayser, C., König, P. & Körding, K. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience*, *15*(3), 475-86.

Feldman, D. E. (2000, July). Timing-based LTP and LTD at vertical input to layer II/III pyramidal cells in rat barrel cortex. *Neuron*, *27*, 45–56.

Földiàk, P. (1989). Adaptive network for optimal linear feature extraction. In *Proceedings of the ieee/inns international joint conference on neural networks* (S. 401-405). New York: IEEE Press.

Földiàk, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Foster, D. J. & Wilson, M. A. (2006, March). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683.

Franzius, M. (2008). *Slowness and sparseness for unsupervised learning of spatial and object codes from naturalistic data.* Unveröffentlichte Dissertation, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät I, Universitätsbibliothek.

Franzius, M., Sprekeler, H. & Wiskott, L. (2007, Aug). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computationl Biology*, *3*(8), e166.

Franzius, M., Vollgraf, R. & Wiskott, L. (2007, Jun). From grids to places. *Journal of Computational Neuroscience*, *22*(3), 297–299.

Franzius, M., Wilbert, N. & Wiskott, L. (2007). Unsupervised learning of invariant 3D-object representations with slow feature analysis. In *Proc. 3rd bernstein symposium for computational neuroscience, göttingen, september 24–27* (S. 105). Göttingen: Bernstein Center for Computational Neuroscience (BCCN).

Friedman, N., Mosenzon, O., Slonim, N. & Tishby, N. (2001). Multivariate information bottleneck. In *Proceedings of uncertainty in ai.* San Francisco, CA, USA: Morgan Kaufman Publishers.

Froemke, R., Merzenich, M. & Schreiner, C. (2007). A synaptic memory trace for cortical receptive field plasticity. *Nature*, *450*(7168), 425–9.

Gardiner, C. W. (1985). *Handbook of stochastic methods for physics, chemistry and the natural sciences* (2nd Aufl.). Berlin & New York: Springer-Verlag.

Gerstner, W., Kempter, R., Hemmen, J. van & Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. *Nature*, *383*(6595), 76–78.

Gerstner, W. & Kistler, W. (2002). *Spiking neuron models.* Cambridge, UK: Cambridge University Press.

Gütig, R., Aharonov, S., Rotter, S. & Sompolinsky, H. (2003). Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. *Journal of Neuroscience*, *23*(9), 3697–3714.

Guyonneau, R., VanRullen, R. & Thorpe, S. (2005). Neurons tune to the earliest spikes through stdp. *Neural Computation*, *17*(4), 859–879.

Hafting, T., Fyhn, M., Molden, S., Moser, M. & Moser, E. I. (2005, August 11). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806.

Harmeling, S., Ziehe, A., Kawanabe, M. & Müller, K.-R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, *15*, 1089–1124.

Hecht, R. M. & Tishby, N. (2005). *Extraction of relevant speech features using the information bottleneck method.*

Hirsch, H. & Spinelli, D. (1971). Modification of the distribution of receptive field orientation in cats by selective visual exposure during development. *Experimental Brain Research*, *12*(5), 509–527.

Horn, D., Levy, N., Meilijson, I. & Ruppin, E. (2000). Distributed synchrony of spiking neurons in a Hebbian cell assembly. In S. A. Solla, T. K. Leen & K.-R. Müller (Hg.), *Advances in neural information processing systems 12* (S. 129–135). New York, NY, USA: MIT Press.

Horton, J. & Adams, D. (2005). The cortical column: a structure without a function. *Philosophical Transactions: Biological Sciences*, *360*(1456), 837–862.

Hosoya, T., Baccus, S. A. & Meister, M. (2005, July). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106–154.

Hubel, D. & Wiesel, T. (1963). Receptive fields of cells in striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, *26*(6), 994–1002.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*, 626–634.

Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*. New York: Wiley.

Hyvärinen, A. & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, *12*(3), 429–439.

Joliffe, L. (1986). *Principal component analysis*. New York: Springer-Verlag.

Jutten, C. & Karhunen, J. (2003). Advances in nonlinear blind source separation. *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 245–256.

Kempter, R., Gerstner, W. & Hemmen, J. L. van. (1999). Hebbian learning and spiking neurons. *Physical Review E*, *59*, 4498–4514.

Kempter, R., Gerstner, W. & Hemmen, J. L. van. (2001). Intrinsic stabilization of output rates by spike-based Hebbian learning. *Neural Computation*, *13*, 2709-2741.

Kepecs, A., Rossum, M. van, Song, S. & Tegner, J. (2002). Spike-timing-dependent plasticity: Common themes and divergent vistas. *Biological Cybernetics*, *87*(5), 446–458.

Kistler, W. M. & Hemmen, J. L. van. (2000). Modeling synaptic plasticity in conjunction with the timing of pre- and postsynaptic action potentials. *Neural Computation*, *12*, 385.

Kittel, C. et al. (1986). *Introduction to solid state physics*. New York: Wiley.

Koch, C., Rapp, M. & Segev, I. (1996). A brief history of time (constants). *Cerebral Cortex*, *6*, 92–101.

Körding, K., Kayser, C., Einhäuser, W. & König, P. (2004). How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, *91*(1), 206–212.

Körding, K. P. & König, P. (2001, December). Neurons with two sites of synaptic integration learn invariant representations. *Neural Computation*, *13*(12), 2823–2849.

Landau, L. D. & Lifshitz, E. M. (1977). *Quantum mechanics: Non-relativistic theory* (Bd. 3). New York: Pergamon Press.

London, M. & Häusser, M. (2005). Dendritic computation. *Annual Reviews of Neuroscience*, *28*, 503–532.

Lu, J., Li, C., Zhao, J., Poo, M. & Zhang, X.(2007). Spike-timing-dependent plasticity of neocortical excitatory synapses on inhibitory interneurons depends on target cell type. *Journal of Neuroscience*, *27*(36), 9711.

Magnus, J. R. & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.

Markram, H., Lübke, J., Frotscher, M. & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, *275*, 213–215.

Markus, E. J., Qin, Y. L., Leonard, B., Skaggs, W. E., McNaughton, B. L. & Barnes, C. A.(1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience*, *15*(11), 7079–7094.

McLaughlin, T. & O'Leary, D.(2005). Molecular gradients and development of retinotopic maps. *Annual Reviews of Neuroscience*, *28*, 327–355.

Mehta, M., Quirk, M. & Wilson, M.(2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*(3), 707–715.

Mehta, M. R., Barnes, C. A. & McNaughton, B. L.(1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *PNAS*, *94*(16), 8918-8921.

Mitchison, G.(1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, *3*, 312–320.

Miyashita, Y.(1988, Oct). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*(6193), 817–820.

Molgedey, L. & Schuster, H. G.(1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, *72*(23), 3634–3637.

Muller, R. U.(1996). A quarter of a century of place cells. *Neuron*, *17*, 979–990.

Muller, R. U., Bostock, E., Taube, J. S. & Kubie, J. L.(1994). On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, *14*(12), 7235–7251.

Murthy, V., Schikorski, T., Stevens, C. & Zhu, Y.(2001). Inactivity Produces Increases in Neurotransmitter Release and Synapse Size. *Neuron*, *32*(4), 673–682.

Oja, E. (1982). A simplified neuron as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267-273.

Oja, E. & Karhunen, J.(1995). Signal separation by nonlinear Hebbian learning. *Computational Intelligence: A Dynamic System Perspective*, 83–97.

O'Keefe, J.(2007). Hippocampal neurophysiology in the behaving animal. In *The hippocampus book* (S. 475–548). Oxford, UK: Oxford university press.

O'Keefe, J. & Dostrovsky, J.(1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research*, *34*, 171–175.

Olshausen, B. A. & Field, D.(2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Olshausen, B. A. & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*, 3311–3325.

O'Reilly, R. C. & Johnson, M. H. (1994). Object recognition and sensitive periods: A computational analysis of visual imprinting. *Neural Computation*, *6*(3), 357–389.

Peng, H. C., Sha, L. F., Gan, Q. & Wei, Y. (1998). Energy function for learning invariance in multilayer perceptron. *Electronics Letters*, *34*(3), 292–294.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Rao, R. P. & Ballard, D. H. (1999, Jan). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rao, R. P. & Sejnowski, T. J. (2001). Spike-timing-dependent Hebbian plasticity as temporal difference learning. *Neural Computation*, *13*(10), 2221–2238.

Reif, F. & Muschik, W. (1987). *Statistische Physik und Theorie der Wärme*. Berlin: de Gruyter.

Riegle, K. C. & Meyer, R. L. (2007). Rapid homeostatic plasticity in the intact adult visual system. *Journal of Neuroscience*, *27*(39), 10556-10567.

Ringach, D. (2007). On the origin of the functional architecture of the cortex. *PLoS ONE*, *2*(2), e251.

Rolls, E. T. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, *9*, 467–480.

Rolls, E. T. (2006). Neurophysiological and computational analyses of the primate presubiculum, subiculum and related areas. *Behavioral Brain Research*, *174*, 289–303.

Rossum, M. C. W. van, Bi, G. & Turrigiano, G. G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience*, *20*(23), 8812–8821.

Rubin, J., Lee, D. D. & Sompolinsky, H. (2001). Equilibrium properties of temporally asymmetric Hebbian learning. *Physical Review Letters*, *86*(2), 364–367.

Ruderman, D. L. & Bialek, W. (1994, Aug). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, *73*(6), 814–817.

Schultz, W. & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Reviews of Neuroscience*, *23*, 472–500.

Sharp, P. E. . (1991). Computer simulation of hippocampal place cells. *Psychobiology*, *19*(2), 103–115.

Shaw, J. (2006). *Unifying perception and curiosity*. Unveröffentlichte Dissertation, University of Rochester.

Sherrington, C. (1906). *The Integrative Action of the Nervous System.* New York: C. Scribner's sons.

Sjöström, P. J., Turrigiano, G. G. & Nelson, S. B. (2001, Dec). Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, *32*(6), 1149-1164.

Slonim, N. & Tishby, N. (2000). *Document clustering using word clusters via the information bottleneck method.* ACM press, New York.

Song, S. & Abbott, L. F. (2001). Cortical mapping and development through spike timing-dependent plasticity. *Neuron*, *32*, 339-350.

Sprekeler, H., Michaelis, C. & Wiskott, L. (2007). Slowness: An objective for spike-timing-plasticity? *PLoS Computational Biology*, *3*(6), e112.

Srinivasan, M. V., Laughlin, S. B. & Dubs, A. (1982, Nov). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *216*(1205), 427–459.

Stone, J. V. & Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, *6*, 429–436.

Sutton, R. & Barto, A. (1998). *Reinforcement learning: An introduction.* New York: MIT Press.

Taube, J. S. & Bassett, J. P. (2003). Persistent neural activity in head direction cells. *Cerebral Cortex*, *13*, 1162–1172.

Taube, J. S., Muller, R. U. & Ranck, J. B. J. (1990). Head direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, *2*(10), 420–435.

Tishby, N., Pereira, F. & Bialek, W. (1999). *The information bottleneck method.*

Torborg, C. & Feller, M. (2005). Spontaneous patterned retinal activity and the refinement of retinal projections. *Progress in Neurobiology*, *76*(4), 312–235.

Turrigiano, G. G. (2007). Homeostatic signaling: the positive side of negative feedback. *Current Opinion in Neurobiology*, *17*, 318–324.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, *391*, 892–895.

Turrigiano, G. G. & Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current Opinion in Neurobiology*, *10*, 358–364.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory.* New York: Springer.

Wallis, G. & Baddeley, R. (1997). Optimal, unsupervised learning in invariant object recognition. *Neural Computation*, *9*, 883-894.

Wallis, G. & Bülthoff, H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, 71028598.

Wallis, G. & Rolls, E. T. (1997, February). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*(2), 167–194.

Wiesel, T. N. (1982, Oct). Postnatal development of the visual cortex and the influence of environment. *Nature*, *299*(5884), 583–591.

Wiskott, L. (2003, September). Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, *15*(9), 2147–2177.

Wiskott, L. & Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770.

Wong, R. (1999). Retinal waves and visual system development. *Annual Review of Neuroscience*, *22*(1), 29–47.

Wyss, R., König, P. & Verschure, P. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, *4*(5), e120.

Yao, H. & Dan, Y. (2001, Oct). Stimulus timing-dependent plasticity in cortical processing of orientation. *Neuron*, *32*(2), 315–323.

Yao, H., Shen, Y. & Dan, Y. (2004). Intracortical mechanism of stimulus-timing-dependent plasticity in visual cortical orientation tuning. *PNAS*, *101*(14), 5081–5086.

Young, J., Waleszczyk, W., Wang, C., Calford, M., Dreher, B. & Obermayer, K. (2007). Cortical reorganization consistent with spike timing–but not correlation-dependent plasticity. *Nature Neuroscience*, *10*, 887–895.

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W. A. & Poo, M. (1998, September). A critical window for cooperation and competition among developing retinotectal synapses. *Nature*, *395*(6697), 37–44.

Ziehe, A. & Müller, K. (1998). TDSEP–an efficient algorithm for blind separation using time structure. *Proc. Int. Conf. on Artificial Neural Networks (ICANN '98)*, 675–680.

Appendix A

Mathematical Derivations

A.1 Proof of Theorem 1 in Chapter 3

Theorem 1. *The solution of optimization problem 2 is given by the J eigenfunctions of the operator $\mathcal{D}$ with the smallest eigenvalues, i.e. the functions that fulfill the eigenvalue equation*

$$\mathcal{D}g_j = \lambda_j g_j \tag{A.1}$$

with the boundary condition

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_i = 0\,. \tag{A.2}$$

Here, the operator $\mathcal{D}$ is given by

$$\mathcal{D} := -\frac{1}{p_{\mathbf{x}}} \sum_{\mu,\nu} \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu \tag{A.3}$$

and the eigenfunctions are assumed to be normalized accoding to

$$(g_j, g_j) = 1\,. \tag{A.4}$$

$\mathbf{n}(\mathbf{x})$ denotes the normal vector on the boundary for the point $\mathbf{x}$. The Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_j) = \lambda_j\,. \tag{A.5}$$

Preliminary Lemmas

For reasons of clarity, we will first prove several lemmas that help to prove Theorem 1. The first lemma shows that the optimal functions for SFA fulfill an Euler-Lagrange equation that is similar to the eigenvalue equation for the operator $\mathcal{D}$.

Lemma 1. *For a particular choice of the parameters λ_{ij}, the solutions g_j of optimization problem 2 obey the Euler-Lagrange equation*

$$\mathcal{D}g_j(\mathbf{x}) - \lambda_{j0} - \lambda_{jj} g_j(\mathbf{x}) - \sum_{i<j} \lambda_{ji} g_i(\mathbf{x}) = 0 \tag{A.6}$$

with the boundary condition (A.2) and the operator $\mathcal{D}$ according to equation (A.3).

Proof. Optimization problem 2 is in essence a constrained optimization problem. The standard technique for such constrained optimization problems is that of Lagrange multipliers. This technique states that the solutions of the optimization problem have to fulfill the necessary condition to be stationary points of an objective function Ψ that incorporates the constraints

$$\Psi(g_j) = \frac{1}{2}\Delta(g_j) - \lambda_{j0}\langle g_j(\mathbf{x})\rangle_{\mathbf{x}} - \frac{1}{2}\lambda_{jj}\langle g_j(\mathbf{x})^2\rangle_{\mathbf{x}} - \sum_{i<j}\lambda_{ji}\langle g_i(\mathbf{x})g_j(\mathbf{x})\rangle_{\mathbf{x}}\,, \tag{A.7}$$

where λ_{ij} are Lagrange multipliers that need to be chosen such that the stationary points fulfill the constraints.

The objective (A.7) is a functional of the function g_j we want to optimize. Because a gradient is difficult to define for functionals, we cannot find the stationary points by simply setting the gradient to zero. Instead, the problem requires variational calculus.

The technique of variational calculus can be illustrated by means of an expansion in the spirit of a Taylor expansion. Let us assume that we know the function g_j that optimizes the objective function Ψ. The effect of a small change δg of g_j on the objective function Ψ can be written as

$$\Psi(g_j + \delta g) - \Psi(g_j) = \int \frac{\delta\Psi}{\delta g_j}(\mathbf{x})\,\delta g(\mathbf{x})\,\mathrm{d}^N x + \ldots\,, \tag{A.8}$$

where the ellipses stand for higher order terms in δg. The function $\frac{\delta\Psi}{\delta g_j}$ is the *variational derivative* of the functional Ψ and usually depends on the input signal $\mathbf{x}$, the optimal function g_j, and possibly derivatives of g_j. Its analogue in finite-dimensional calculus is the gradient.

We now derive an expression for the variational derivative of the objective function (A.7). To keep the calculations tidy, we split the objective in two parts and omit the dependence on the input signal $\mathbf{x}$:

$$\Psi(g_j) =: \frac{1}{2}\Delta(g_j) - \tilde{\Psi}(g_j)\,. \tag{A.9}$$

The expansion of $\tilde{\Psi}$ is straightforward:

$$\tilde{\Psi}(g_j + \delta g) - \tilde{\Psi}(g_j) = \langle \delta g\,[\lambda_{j0} + \lambda_{jj}g_j + \sum_{i<j}\lambda_{ji}g_i]\rangle_{\mathbf{x}} + \ldots \tag{A.10}$$

$$= \int \delta g\, p_{\mathbf{x}}\,[\lambda_{j0} + \lambda_{jj}g_j + \sum_{i<j}\lambda_{ji}g_i]\mathrm{d}^N x + \ldots \tag{A.11}$$

The expansion of $\Delta(g_j)$ is done after expressing the Δ-value in terms of probability density $p_{\mathbf{x}}$ and the matrix $K_{\mu\nu}$ (cf. equation (3.12)):

$$\frac{1}{2}\left[\Delta(g_j + \delta g) - \Delta(g_j)\right] \overset{(3.12)}{=} \frac{1}{2}\sum_{\mu,\nu}\langle K_{\mu\nu}[\partial_\mu(g_j + \delta g)][\partial_\nu(g_j + \delta g)]\rangle_{\mathbf{x}} \tag{A.12}$$

$$- \frac{1}{2}\sum_{\mu,\nu}\langle K_{\mu\nu}[\partial_\mu g_j][\partial_\nu g_j]\rangle_{\mathbf{x}} \tag{A.13}$$

$$= \frac{1}{2}\sum_{\mu,\nu}\langle K_{\mu\nu}[\partial_\mu g_j][\partial_\nu \delta g] + K_{\mu\nu}[\partial_\mu \delta g][\partial_\nu g_j]\rangle_{\mathbf{x}} + \ldots \tag{A.14}$$

$$= \sum_{\mu,\nu} \langle K_{\mu\nu} [\partial_\mu \delta g][\partial_\nu g_j] \rangle_{\mathbf{x}} + ... \tag{A.15}$$

(since $K_{\mu\nu}$ is symmetric)

$$\overset{(3.4)}{=} \sum_{\mu,\nu} \int p_{\mathbf{x}} K_{\mu\nu} [\partial_\mu \delta g][\partial_\nu g_j] \, \mathrm{d}^N x \tag{A.16}$$

$$= \sum_{\mu,\nu} \int \partial_\mu \delta g \, p_{\mathbf{x}} \, K_{\mu\nu} \partial_\nu g_j \, \mathrm{d}^N x \tag{A.17}$$

$$- \sum_{\mu,\nu} \int \delta g \, \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j \, \mathrm{d}^N x + ... \tag{A.18}$$

$$= \sum_{\mu,\nu} \int_{\partial V} \delta g \, p_{\mathbf{x}} \, n_\mu K_{\mu\nu} \partial_\nu g_j \, \mathrm{d}A \tag{A.19}$$

$$- \sum_{\mu,\nu} \int \delta g \, \partial_\mu \left[p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j \right] \, \mathrm{d}^N x + ... \tag{A.20}$$

(Gauss' theorem)

$$\overset{(A.3)}{=} \int_{\partial V} \delta g \sum_{\mu,\nu} n_\mu \, p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_j \, \mathrm{d}A \tag{A.21}$$

$$+ \int \delta g \, p_{\mathbf{x}} \, (\mathcal{D} g_j) \, \mathrm{d}^N x + ... \, . \tag{A.22}$$

Here, $\mathrm{d}A$ is an infinitesimal surface element of the boundary ∂V of V and $\mathbf{n}$ is the normal vector on $\mathrm{d}A$. To get the expansion of the full objective function, we add (A.11) and (A.22):

$$\begin{aligned} \Psi(g_j + \delta g) - \Psi(g_j) = & \int_{\partial V} \delta g \sum_{\mu,\nu} n_\mu \, p_{\mathbf{x}} K_{\mu\nu} \partial_\mu g_j \, \mathrm{d}A \\ & + \int \delta g \, p_{\mathbf{x}} \, (\mathcal{D} g_j - \lambda_{j0} - \lambda_{jj} g_j - \sum_{i<j} \lambda_{ji} g_i) \, \mathrm{d}^N x + ... \, . \end{aligned} \tag{A.23}$$

In analogy to the finite-dimensional case, g_j can only be an optimum of the objective function Ψ if any small change δg leaves the objective unchanged up to linear order. As we employ a Lagrange multiplier ansatz, we have an unrestricted optimization problem, so we are free in choosing δg. From this it is clear that the right hand side of (A.23) can only vanish if the integrands of both the boundary and the volume integral vanish separately. This leaves us with the differential equation (A.6) and the boundary condition (A.2). □

Next, we show that the operator $\mathcal{D}$ is self-adjoint with respect to the scalar product (3.13) when restricted to the set of functions that fulfill the boundary condition (A.2).

Lemma 2. *Let $\mathcal{F}_b \subset \mathcal{F}$ be the space of functions obeying the boundary condition (3.24,A.2). Then $\mathcal{D}$ is self-adjoint on $\mathcal{F}_b$ with respect to the scalar product*

$$(f, g) := \langle f(\mathbf{x}) g(\mathbf{x}) \rangle_{\mathbf{x}}, \tag{A.24}$$

i.e.

$$\forall f, g \in \mathcal{F}_b : (\mathcal{D} f, g) = (f, \mathcal{D} g). \tag{A.25}$$

Proof. The proof can be carried out in a direct fashion. Again, we omit the explicit dependence on $\mathbf{x}$.

$$
\begin{aligned}
(f, \mathcal{D}g) &\overset{(A.24,A.3,3.4)}{=} -\int p_{\mathbf{x}} f \frac{1}{p_{\mathbf{x}}} \sum_{\mu,\nu} \partial_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g \, \mathrm{d}^N x && \text{(A.26)} \\
&= -\sum_{\mu,\nu} \int \partial_\mu p_{\mathbf{x}} f K_{\mu\nu} \partial_\nu g \, \mathrm{d}^N x + \int p_{\mathbf{x}} \sum_{\mu,\nu} K_{\mu\nu} [\partial_\mu f][\partial_\nu g] \, \mathrm{d}^N x \\
&= -\int_{\partial V} f \underbrace{\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g}_{\overset{(A.2)}{=}0} \, \mathrm{d}A + \int p_{\mathbf{x}} \sum_{\mu,\nu} K_{\mu\nu} [\partial_\mu f][\partial_\nu g] \mathrm{d}^N x \\
&\quad \text{(Gauss' theorem)} && \text{(A.27)} \\
&\overset{(A.2)}{=} \int p_{\mathbf{x}} K_{\mu\nu} [\partial_\mu f][\partial_\nu g] \, \mathrm{d}^N x && \text{(A.28)} \\
&= \int p_{\mathbf{x}} K_{\mu\nu} [\partial_\mu g][\partial_\nu f] \, \mathrm{d}^N x && \text{(A.29)} \\
&\quad \text{(since } K_{\mu\nu} \text{ is symmetric)} \\
&\overset{(A.26-A.28)}{=} (\mathcal{D}f, g) \,. && \text{(A.30)}
\end{aligned}
$$

□

This property is useful, because it allows the application of the spectral theorem known from functional analysis (Courant & Hilbert, 1989), which states that any self-adjoint operator possesses a complete set of eigenfunctions $f_j(\mathbf{s}) \in \mathcal{F}_b$ with real eigenvalues Δ_j, which are pairwise orthogonal, i.e. a set of functions that fulfill the following conditions:

$$
\begin{aligned}
\mathcal{D}f_j &= \Delta_j f_j \quad \text{with } \Delta_j \in \mathbb{R} && \text{(eigenvalue equation)} \,, && \text{(A.31)} \\
(f_i, f_j) &= \delta_{ij} && \text{(orthonormality)} \,, && \text{(A.32)} \\
\forall f \in \mathcal{F}_b \, \exists \, \alpha_k : f &= \sum_{k=0}^{\infty} \alpha_k f_k && \text{(completeness)} \,. && \text{(A.33)}
\end{aligned}
$$

The eigenfunctions, normalized according to (A.32), thus fulfill the unit variance and decorrelation constraints (3.16). If we set $\lambda_{0j} = \lambda_{ji} = 0$ for $i \neq j$, the eigenfunctions also solve the Euler-Lagrange equation (A.6), which makes them good candidates for the solution of optimization problem 2. To show that they indeed minimize the Δ-value we need

Lemma 3. *The Δ-value of the normalized eigenfunctions f_j is given by their eigenvalue Δ_j.*

Proof.

$$
\Delta(f_j) \overset{(3.12,3.4,A.26-A.28)}{=} (f_j, \mathcal{D}f_j) \overset{(A.31)}{=} (f_j, \Delta_j f_j) = \Delta_j \underbrace{(f_j, f_j)}_{=1} \overset{(A.32)}{=} \Delta_j \,. \qquad \text{(A.34)}
$$

□

Proof of Theorem 1

At this point, we have everything we need to prove

Theorem 1. *The solution of optimization problem 2 is given by the J eigenfunctions of the operator $\mathcal{D}$ with the smallest eigenvalues, i.e. the functions that fulfill*

$$\mathcal{D}g_i = \lambda_i g_i \tag{A.35}$$

with the boundary condition

$$\sum_{\mu,\nu} n_\mu p_{\mathbf{x}} K_{\mu\nu} \partial_\nu g_i = 0 \tag{A.36}$$

and the normalization condition

$$(g_i, g_i) = 1\,. \tag{A.37}$$

The Δ-value of the eigenfunctions is given by their eigenvalue

$$\Delta(g_i) = \lambda_i\,. \tag{A.38}$$

Proof. Without loss of generality we assume that the eigenfunctions f_j are ordered by increasing eigenvalue, starting with the constant $f_0 = 1$. There are no negative eigenvalues, because according to Lemma 3, the eigenvalue is the Δ-value of the eigenfunction, which can only be positive by definition. According to Lemma 1, the optimal responses g_j obey the boundary condition (A.2) and are thus elements of the subspace $\mathcal{F}_b \subset \mathcal{F}$ defined in Lemma 2. Because of the completeness of the eigenfunctions on $\mathcal{F}_b$ we can do the expansion

$$g_j = \sum_{k=1}^{\infty} \alpha_{jk} f_k \tag{A.39}$$

where we may omit f_0 because of the zero mean constraint. We can now prove by complete induction that $g_j = f_j$ solves the optimization problem.
Basis (j=1): Inserting g_1 into equation (A.6) we find

$$0 = \mathcal{D}g_1 - \lambda_{10} - \lambda_{11} g_1 \tag{A.40}$$

$$\overset{(A.39,A.31)}{=} -\lambda_{10} + \sum_{k=1}^{\infty} \alpha_{1k}(\Delta_k - \lambda_{11}) f_k \tag{A.41}$$

$$\Rightarrow \quad \begin{array}{c} \lambda_{10} = 0 \\ \wedge \quad (\alpha_{1k} = 0 \vee \Delta_k = \lambda_{11})\,\forall k\,, \end{array} \tag{A.42}$$

because f_k and the constant are linearly independent and (A.40) must be fulfilled for all $\mathbf{x}$. Equation (A.42) implies that the coefficients α_{1k} have to vanish unless the Δ-value of the associated eigenfunction is equal to λ_{11}. Thus, only eigenfunctions that have the same Δ-value can have non-vanishing coefficients. Therefore, the optimal response g_1 must also be an eigenfunction of $\mathcal{D}$. Since the Δ-value of the eigenfunctions is given by their eigenvalue, it is obviously optimal to chose $g_1 = f_1$. Note that although this choice is optimal, it is not necessarily unique, since there may be several eigenfunctions with the same eigenvalue. In this case any linear combination of these functions is also optimal.
Induction step: Given that $g_i = f_i$ for $i < j$, we prove that $g_j = f_j$ is optimal. Because of the orthonormality of the eigenfunctions the decorrelation constraint (3.16) yields

$$0 \overset{(3.16)}{=} (g_i, g_j) = (f_i, \sum_{k=1}^{\infty} \alpha_{jk} f_k) = \alpha_{ji} \quad \forall i < j\,. \tag{A.43}$$

Again inserting the expansion (A.39) into (A.6) yields

$$0 \overset{(A.6,A.39)}{=} (\mathcal{D} - \lambda_{jj}) \sum_{k=1}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i \tag{A.44}$$

$$\overset{(A.43)}{=} (\mathcal{D} - \lambda_{jj}) \sum_{k=j}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i \tag{A.45}$$

$$\overset{(A.31)}{=} \sum_{k=j}^{\infty} \alpha_{jk} (\Delta_k - \lambda_{jj}) f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i \tag{A.46}$$

$$\begin{aligned} \Longrightarrow \quad & \lambda_{j0} = 0 \\ \wedge \quad & \lambda_{ji} = 0 \qquad \forall i < j \\ \wedge \quad & (\alpha_{jk} = 0 \vee \Delta_k = \lambda_{jj}) \quad \forall k \geq j \,, \end{aligned} \tag{A.47}$$

because the eigenfunctions f_i are linearly independent. The conditions (A.47) can only be fulfilled if g_j is an eigenfunction of $\mathcal{D}$. Because of Lemma 3 an optimal choice for minimizing the Δ-value without violating the decorrelation constraint is $g_j = f_j$.

□

A.2 Derivation of the Generators Used in Chapter 5

transformation	**generator**	**velocity**
translation	$\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'}$	$\mathbf{v}$
rotation	$r_1 \partial_{r_2} - r_2 \partial_{r_1} + r'_1 \partial_{r'_2} - r'_2 \partial_{r'_1}$	ω
zoom	$\mathbf{r} \cdot \nabla_{\mathbf{r}} + \mathbf{r}' \cdot \nabla_{\mathbf{r}'} + 4$	$\zeta = \dot{z}/z$

Table A.1: Generators of the transformations used to generate the image sequences. ∂_{r_1} denotes the derivative with respect to the first component of $\mathbf{r}$. $\nabla_{\mathbf{r}}$ denotes the vector-valued operator $(\partial_{r_1}, \partial_{r_2})^T$. This table is identical with table 5.2 and only repeated for convenience.

Translation

We use the convention that the effect of a translation T_x of an image $x(\mathbf{r})$ by a vector $\mathbf{R}$ is the replacement of the pixel value at position $\mathbf{r}$ by the pixel value of the original image at the position $\mathbf{r} - \mathbf{R}$:

$$(T_x x)(\mathbf{r}) = x(\mathbf{r} - \mathbf{R}) \,. \tag{A.48}$$

What is the corresponding representation of translation on the quadratic functions (5.1)? This can be seen immediately by means of a variable substitution:

$$(T_g g)[x(\mathbf{r})] \overset{(5.3)}{:=} \int \tilde{g}(\mathbf{r}, \mathbf{r}') (T_x x)(\mathbf{r}) (T_x x)(\mathbf{r}') \, \mathrm{d}^2 r \, \mathrm{d}^2 r' \tag{A.49}$$

$$\overset{(A.48)}{=} \int \tilde{g}(\mathbf{r}, \mathbf{r}') x(\mathbf{r} - \mathbf{R}) x(\mathbf{r}' - \mathbf{R}) \, \mathrm{d}^2 r \, \mathrm{d}^2 r' \tag{A.50}$$

$$= \int \tilde{g}(\mathbf{r} + \mathbf{R}, \mathbf{r}' + \mathbf{R}) x(\mathbf{r}) x(\mathbf{r}') \, \mathrm{d}^2 r \, \mathrm{d}^2 r' \,. \tag{A.51}$$

Thus, the effect of the translation operator on the functional g is the replacement of the kernel $\tilde{g}(\mathbf{r}, \mathbf{r}')$ by $\tilde{g}(\mathbf{r} + \mathbf{R}, \mathbf{r}' + \mathbf{R})$:

$$(T_g \tilde{g})(\mathbf{r}, \mathbf{r}') = \tilde{g}(\mathbf{r} + \mathbf{R}, \mathbf{r}' + \mathbf{R})\,. \tag{A.52}$$

Remember that we represent the functionals in terms of the basis functions $x(\mathbf{r})x(\mathbf{r}')$. In this basis, the functional g is represented by the "coefficient function" $\tilde{g}(\mathbf{r}, \mathbf{r}')$. Equation (A.52) is the representation of the translation operator in this basis.

We can now calculate the associated generator by applying a time-dependent translation $T_g(t)$ by a vector $\mathbf{R}(t)$ and calculating the temporal derivative:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(T_g(t)\tilde{g})(\mathbf{r}, \mathbf{r}') &\overset{(A.52)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}(\mathbf{r} + \mathbf{R}(t), \mathbf{r}' + \mathbf{R}(t)) && \text{(A.53)}\\
&= \left(\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{R}(t) \cdot [\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'}]\,\tilde{g}\right)(\mathbf{r} + \mathbf{R}(t), \mathbf{r}' + \mathbf{R}(t)) && \text{(A.54)}\\
&= (T_g(t)\, Q^{\mathrm{trans}}(t)\tilde{g})(\mathbf{r}, \mathbf{r}') && \text{(A.55)}
\end{aligned}$$

with

$$Q^{\mathrm{trans}}(t) := \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{R}(t) \cdot [\nabla_{\mathbf{r}} + \nabla_{\mathbf{r}'}]\,. \tag{A.56}$$

Clearly, the translation velocity $\mathbf{v} := \mathrm{d}\mathbf{R}/\mathrm{d}t$ plays the role of the velocity in equation (5.10), while the sum of the gradients is the generator of translations as stated in table 5.2.

Rotation

A rotation of an image $x(\mathbf{r})$ by an angle ϕ corresponds to the application of an orthogonal matrix $\mathbf{O}^{-1} = \mathbf{O}^T$ to the pixel positions:

$$(T_x x)(\mathbf{r}) = x(\mathbf{O}^T\mathbf{r})\,, \tag{A.57}$$

where

$$\mathbf{O} = \begin{pmatrix} \cos(\phi) & \sin(\phi) \\ \sin(-\phi) & \cos(\phi) \end{pmatrix}. \tag{A.58}$$

The effect of the related rotation operator T_g on the integral kernel $g(\mathbf{r}, \mathbf{r}')$ can again be derived by a variable substitution:

$$\begin{aligned}
(T_g g)[x(\mathbf{r})] &\overset{(5.3)}{:=} \int \tilde{g}(\mathbf{r}, \mathbf{r}')(T_x x)(\mathbf{r})(T_x x)(\mathbf{r}')\,\mathrm{d}^2 r\,\mathrm{d}^2 r' && \text{(A.59)}\\
&\overset{(A.57)}{=} \int \tilde{g}(\mathbf{r}, \mathbf{r}')x(\mathbf{O}^T\mathbf{r})x(\mathbf{O}^T\mathbf{r}')\,\mathrm{d}^2 r\,\mathrm{d}^2 r' && \text{(A.60)}\\
&= \int \tilde{g}(\mathbf{O}\mathbf{r}, \mathbf{O}\mathbf{r}')x(\mathbf{r})x(\mathbf{r}')\,\mathrm{d}^2 r\,\mathrm{d}^2 r'\,. && \text{(A.61)}
\end{aligned}$$

Thus, in the basis $x(\mathbf{r})x(\mathbf{r})'$, rotations are represented by

$$(T_g \tilde{g})(\mathbf{r}, \mathbf{r}') = \tilde{g}(\mathbf{O}\mathbf{r}, \mathbf{O}\mathbf{r}')\,. \tag{A.62}$$

Again, we can calculate the generator by taking the temporal derivative of a time-dependent rotation $T_g(t)$ by a matrix $\mathbf{O}(t)$. To keep the notation short, we omit the

time dependence of the rotation matrix $\mathbf{O}$ and use the short notation $\dot{\mathbf{O}} := \frac{\mathrm{d}\mathbf{O}}{\mathrm{d}t}$ for its temporal derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}(T_g(t)\tilde{g})(\mathbf{r},\mathbf{r}') \overset{(A.62)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}(\mathbf{Or},\mathbf{Or}') \tag{A.63}$$

$$= \left(\left[(\dot{\mathbf{O}}\mathbf{r})\cdot\nabla_{\mathbf{r}} + (\dot{\mathbf{O}}\mathbf{r}')\cdot\nabla_{\mathbf{r}'}\right]\tilde{g}\right)(\mathbf{Or},\mathbf{Or}') \tag{A.64}$$

$$= \Big(T_g(t)\underbrace{\left[(\dot{\mathbf{O}}\mathbf{O}^T\mathbf{r})\cdot\nabla_{\mathbf{r}} + (\dot{\mathbf{O}}\mathbf{O}^T\mathbf{r}')\cdot\nabla_{\mathbf{r}'}\right]}_{=:Q^{\mathrm{rot}}(t)}\tilde{g}\Big)(\mathbf{r},\mathbf{r}') \tag{A.65}$$

$$= \left(T_g(t)\,Q^{\mathrm{rot}}(t)\tilde{g}\right)(\mathbf{r},\mathbf{r}')\,. \tag{A.66}$$

The matrix $\dot{\mathbf{O}}\mathbf{O}^T$ is antisymmetric, because

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{I} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{O}\mathbf{O}^T) = \dot{\mathbf{O}}\mathbf{O}^T + \mathbf{O}\dot{\mathbf{O}}^T = \dot{\mathbf{O}}\mathbf{O}^T + (\dot{\mathbf{O}}\mathbf{O}^T)^T\,. \tag{A.67}$$

Here, $\mathbf{I}$ denotes the unit matrix. Because of the antisymmetry, $\dot{\mathbf{O}}\mathbf{O}^T$ can be written as

$$\dot{\mathbf{O}}\mathbf{O}^T = \begin{pmatrix} 0 & -\omega(t) \\ \omega(t) & 0 \end{pmatrix}. \tag{A.68}$$

It can be shown that $\omega(t) = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$ is the angular velocity of the rotation. In this notation, $Q^{\mathrm{rot}}(t)$ becomes

$$Q^{\mathrm{rot}}(t) = \omega(t)\left[r_1\partial_{r_2} - r_2\partial_{r_1} + r_1'\partial_{r_2'} - r_2'\partial_{r_1'}\right], \tag{A.69}$$

which leaves us with the generator and the associated velocity ω given in table 5.2.

Zoom

Zooming an image by a zoom factor z around the origin corresponds to replacing the pixel value at position $\mathbf{r}$ by the pixel value of the original image at position $\mathbf{r}/z$. Using similar considerations as above, this leads to the following representation of the zoom operator:

$$(T_g\tilde{g})(\mathbf{r},\mathbf{r}') = z^4\tilde{g}(z\mathbf{r},z\mathbf{r}')\,. \tag{A.70}$$

The factor z^4 is the Jacobian determinant that arises from the coordinate changes $\mathbf{r}\to\mathbf{r}/z$ and $\mathbf{r}'\to\mathbf{r}'/z$ in the integration for $g[x(\mathbf{r})]$.

The generator can again be calculated by introducing a time-dependent zoom factor $z(t)$ and taking the temporal derivative:

$$\frac{\mathrm{d}}{\mathrm{d}t}(T_g(t)\tilde{g})(\mathbf{r},\mathbf{r}') = \frac{\mathrm{d}}{\mathrm{d}t}z^4\tilde{g}(z\mathbf{r},z\mathbf{r}') \tag{A.71}$$

$$= \left(z^4\left[\dot{z}\mathbf{r}\cdot\nabla_{\mathbf{r}} + \dot{z}\mathbf{r}'\cdot\nabla_{\mathbf{r}'}\right] + 4z^3\dot{z}\right)\tilde{g}(z\mathbf{r},z\mathbf{r}') \tag{A.72}$$

$$= z^4\frac{\dot{z}}{z}\left(\left[(z\mathbf{r})\cdot\nabla_{\mathbf{r}} + (z\mathbf{r}')\cdot\nabla_{\mathbf{r}'}\right] + 4\right)\tilde{g}(z\mathbf{r},z\mathbf{r}') \tag{A.73}$$

$$= \left(T_g(t)\,Q^{\mathrm{zoom}}(t)\tilde{g}\right)(\mathbf{r},\mathbf{r}')\,, \tag{A.74}$$

with an operator $Q^{\mathrm{zoom}}(t)$ that contains the generator and the velocity $\zeta := \frac{\dot{z}}{z}$ for zoom as given in table 5.2:

$$Q^{\mathrm{zoom}}(t) = \frac{\dot{z}}{z}\left[\mathbf{r}\cdot\nabla_{\mathbf{r}} + \mathbf{r}'\cdot\nabla_{\mathbf{r}'} + 4\right]. \tag{A.75}$$

A.3 Solution of the Gaussian Information Bottleneck

We start with the objective function (6.15) for the Gaussian information bottleneck:

$$\mathcal{L} = \frac{1-\beta}{2} \ln |\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I}| + \frac{\beta}{2} \ln |\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I}| \,. \tag{A.76}$$

The derivative of the objective function with respect to the weight matrix is given by

$$\frac{d\mathcal{L}}{d\mathbf{A}} = (1-\beta)(\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I})^{-1}\mathbf{A}\mathbf{C}_\mathbf{x} + \beta(\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I})^{-1}\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}} \,. \tag{A.77}$$

By equating to zero and rearranging, we obtain a necessary condition for the weight matrix $\mathbf{A}$:

$$\frac{\beta-1}{\beta} \underbrace{(\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I})(\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I})^{-1}}_{=:\mathbf{M}} \mathbf{A} = \mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{C}_\mathbf{x}^{-1} \,. \tag{A.78}$$

The goal is to prove that this equation can be solved by filling the rows of $\mathbf{A}$ with adequately scaled versions of the solutions $\mathbf{w}_j^T$ of the following generalized eigenvalue problem:

$$\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{w}_j = \lambda_j \mathbf{C}_\mathbf{x}\mathbf{w}_j \,. \tag{A.79}$$

It is always possible to choose the eigenvectors $\mathbf{w}_j$ such that they are orthonormal with respect to the scalar product induced by $\mathbf{C}_\mathbf{x}$:

$$\mathbf{w}_i^T \mathbf{C}_\mathbf{x} \mathbf{w}_j = \delta_{ij} \,. \tag{A.80}$$

By inserting the eigenvectors of the generalized eigenvalue problem into equation (A.78), we will first show that this yields $\mathbf{M}$ diagonal. It then becomes clear that there are scaling factors for the eigenvectors such that equation (A.78) is solved.

(1) **$\mathbf{M}$ is diagonal:** Assume that the rows of $\mathbf{A}$ are filled with the eigenvectors $\mathbf{w}_j^T$, scaled by a factor α_j. Because of the orthonormality condition (A.80), $\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^T + \mathbf{I}$ is then diagonal with diagonal elements $\alpha_j^2 + 1$. Left multiplication of (A.79) with $\mathbf{w}_j^T$ yields that $\mathbf{A}\mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{A}^T + \mathbf{I}$ is also diagonal with diagonal elements $\lambda_j\alpha_j^2 + 1$. Thus, $\mathbf{M}$ is diagonal with diagonal elements

$$M_{jj} = \frac{\lambda_j\alpha_j^2 + 1}{\alpha_j^2 + 1} \,. \tag{A.81}$$

(2) **Scaling factors and critical values for the trade-off parameter β:** The right hand side of equation (A.78) also takes a simple form if $\mathbf{A}$ contains the eigenvectors $\mathbf{w}_j$:

$$\alpha_j \mathbf{w}_j^T \mathbf{C}_{\mathbf{x}|\mathbf{r}}\mathbf{C}_\mathbf{x}^{-1} \overset{(A.79)}{=} \alpha_j\lambda_j \mathbf{w}_j^T \mathbf{C}_\mathbf{x}\mathbf{C}_\mathbf{x}^{-1} = \lambda_j\alpha_j\mathbf{w}_j^T \,. \tag{A.82}$$

Using the diagonal structure (A.81) of $\mathbf{M}$, equation (A.78) becomes

$$\left[\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2+1}{\alpha_j^2+1} - \lambda_j\right] \alpha_j\mathbf{w}_j^T = \mathbf{0} \quad \forall j \,. \tag{A.83}$$

This equation can only be solved by either $\alpha_j = 0$ or by

$$\frac{\beta-1}{\beta}\frac{\lambda_j\alpha_j^2+1}{\alpha_j^2+1} = \lambda_j \,. \tag{A.84}$$

Rearranging for α_j^2 yields the scaling factors α_j:

$$\alpha_j^2 = \frac{\beta(1-\lambda_j)-1}{\lambda_j}\,. \tag{A.85}$$

Of course this equation can only be fulfilled if the right hand side is non-negative. Because λ_j is positive, this reduces to a relation between the β-value and the eigenvalues:

$$\beta \geq \frac{1}{1-\lambda_j}\,. \tag{A.86}$$

For the eigenvalues that do not fulfill this condition for a given β, equation (A.83) can only be solved by $\alpha_j = 0$. This shows that the critical β-values

$$\beta_j^c = \frac{1}{1-\lambda_j} \tag{A.87}$$

as stated in section 6.2 are those, where a new eigenvector becomes available. Moreover, we have now demonstrated that equation (6.16) for $\mathbf{A}(\beta)$ is a solution of equation (A.78) and thus a stationary point of the objective function (6.15) of the Gaussian information bottleneck. In line with the fact that the objective function (6.15) of the Gaussian information bottleneck is invariant with respect to orthogonal transformations of the output signals, it can be shown that any matrix $\tilde{\mathbf{A}} = \mathbf{O}\mathbf{A}$ with $\mathbf{O}^{-1} = \mathbf{O}^T$ is also a solution of (A.78).

At this point we have derived a set of stationary points of the objective function: For all eigenvectors whose eigenvalues fulfill the condition (A.86), the coefficient α_j can either take a finite value according to equation (A.85) or vanish. Intuitively, it is clear that choosing $\alpha_j = 0$ without necessity leads to a loss of information about the input, so that this choice is unlikely to optimize the objective function. For a proof that for optimal choice for the coefficients α_j is indeed given by (A.85), we refer the reader to (Chechik et al., 2005).

Deutsche Zusammenfassung

Einleitung

Das Problem der invarianten Objekterkennung

Die Verarbeitung von Sinneseindrücken durch unser Nervensystem ist eines der zentralen Themen der Neurowissenschaften. Für uns alle ist selbstverständlich, dass unsere Umwelt sich in Objekte zerlegen lässt, aber wie unser Gehirn das visuelle Abbild, das die Welt auf der Netzhaut erzeugt, in Objekte zergliedert, ist weitgehend unverstanden. Sich im Stimmengewirr einer Feier auf Einzelstimmen zu konzentrieren, ist für eine reibungslose Kommunikation unabdingbar, aber die Filtermechanismen, die das Gehirn dafür verwendet, sind unbekannt. Bei genauerer Betrachtung erweisen sich viele Probleme, die unser Gehirn im Zusammenhang mit der Verarbeitung unserer Sinneseindrücke tagtäglich lösen muss, als verblüffend komplex.

Die Motivation der vorliegenden Arbeit ist die Frage, wie wir Objekte wiedererkennen, obwohl die Eindrücke, die unsere Sinnesorgane empfangen, niemals die gleichen sind. Jeder Fotograf weiß, dass das visuelle Bild, das ein Objekt auf unserer Netzhaut/dem Film hinterlässt, von den Lichtverhältnissen, von der Position des Objektes im Sichtfeld und von seiner Orientierung im Raum abhängt. Die Liste der Faktoren, die Einfluss auf den primären Sinneseindruck haben, lässt sich beliebig fortführen, so dass ein Objekt in Abhängigkeit des Kontextes eine schwindelerregende Anzahl verschiedener Sinneseindrücke erzeugen kann. Nichtsdestotrotz sind wir in der Lage, Objekte unabhängig von diesen Faktoren zuverlässig zu erkennen. Welche Mechanismen unser Gehirn dafür verwendet, ist noch unklar.

Ein faszinierender Aspekt dieser Fähigkeit zur *invarianten Objekterkennung* ist, dass sie wahrscheinlich nicht angeboren ist, sondern zumindest teilweise im Laufe der persönlichen Entwicklung erlernt wird. Welche Mechanismen könnten diesem Lernprozess zu Grunde liegen? Welche Indizien könnten unserem Nervensystem einen Hinweis darauf geben, dass zwei visuelle Eindrücke auf ein und dasselbe Objekt zurückzuführen sind?

Das Langsamkeitsprinzip

Eine Beobachtung, die einigen Studien zur invarianten Objekterkennung zu Grunde liegt, ist, dass Objekte typischerweise eine gewisse Zeit in unserem Umfeld verweilen. Deshalb enthalten Sinneseindrücke, die in kurzer zeitlicher Abfolge auftreten, mit großer Wahrscheinlichkeit dieselben Objekte. Dementsprechend sollten sich Signale in unserem Gehirn, die die Präsenz eines Objektes kodieren, auf ebendieser Zeitskala verändern. Im Gegensatz dazu verändern sich die primären Eindrücke, die unsere Sinnesorgane empfangen, häufig auf sehr viel kürzeren Zeitskalen. Zum Beispiel ist der Bereich des Sicht-

feldes, der einer einzelnen retinalen Rezeptorzelle zugänglich ist, sehr klein, so dass sich der Inhalt des relevanten Bildausschnitts schon bei minimalen Veränderungen der Blickrichtung völlig verändern kann. Es ist bekannt, das unser Blickwinkel ständig kleinen, aber schnellen Schwankungen unterlegen ist. Deshalb werden die Signale, die retinale Rezeptorneurone empfangen, vermutlich sehr schnell variieren.

Die Beobachtung, dass sich verhaltensrelevante Signale wie Objektidentitäten im Mittel langsamer verändern als primäre Sinneseindrücke, ist die Basis des so genannten *Langsamkeitsprinzips*. Die Idee ist, dass interne Parameter im Gehirn (oder allgemeiner, in einem lernfähigen System) sich derart anpassen, dass eine Repräsentation der Sinneseindrücke (oder allgemeiner, der Eingangsdaten) entsteht, die sich im Mittel möglichst langsam verändert. Auf diese Weise erhöht sich die Wahrscheinlichkeit, dass die erzeugten Signale verhaltensrelevante Aspekte der Umwelt repräsentieren. Zudem kann sich die gelernte Repräsentation nur dann langsam verändern, wenn sie invariant gegenüber schnell erfolgenden Veränderungen der Umwelt ist. Deshalb ist das Langsamkeitsprinzip ein vielversprechender Kandidat für das Lernen von invarianten Objektdarstellungen.

Dieses Lernprinzip bildet die Grundlage einer ganzen Klasse von Lernalgorithmen. In dieser Arbeit liegt das Augenmerk vornehmlich auf einem von Wiskott und Sejnowski eingeführten Algorithmus, der Slow Feature Analysis (SFA; Wiskott und Sejnowski, 2002).

Slow Feature Analysis

Im Falle von SFA wird das Langsamkeitsprinzip als abstraktes Optimierungsproblem formuliert, das sich durch einen effizienten Algorithmus lösen lässt. Dabei werden die Eingangsdaten (z.B. primäre sensorische Signale) durch einen Satz von Funktionen (die z.B. die Signalverarbeitung durch das Gehirn darstellen sollen) auf eine Repräsentation abgebildet, die sich zeitlich langsam verändert. Eine wichtige Einschränkung ist hierbei, dass die Funktionen die Eingangsdaten *instantan* verarbeiten, so dass trotz der Zielsetzung, langsame Signale zu erzeugen, keinerlei Einschränkung der Verarbeitungsgeschwindigkeit entsteht. Deshalb können langsame Signale nur dann aus den Eingangsdaten extrahiert werden, wenn diese Aspekte enthalten, die sich langsam verändern. Die Komplexität der Funktionen wird zudem beschränkt, indem man einen Funktionenraum festlegt, aus dem sie stammen müssen. Die Aufgabe von SFA ist, diejenigen Funktionen im Funktionenraum zu finden, die für einen gegebenen Satz von *Trainingsdaten* die langsamsten Ausgangssignale erzeugen. Die Langsamkeit der Ausgangssignale wird dabei durch den Mittelwert des Quadrats ihrer Zeitableitung, den so genannten Δ-Wert, quantifiziert. Kleine Δ-Werte entsprechen langsamen Signalen. Die erste Funktion, die SFA findet, ist diejenige im Funktionenraum, deren Ausgangssignal den kleinsten Δ-Wert hat. Die zweite erzeugt das langsamste Ausgangssignal, das sich unter der Bedingung, dass es unkorreliert zum Ausgangssignal der ersten Funktion ist, erzeugen lässt. Das dritte Ausgangssignal wiederum soll unkorreliert zu den ersten beiden sein. Auf diese Weise wird eine Reihe von Funktionen ausgewählt, die unkorrelierte Signale erzeugen und nach ihrer Langsamkeit sortiert sind. Eine detaillierte Darstellung von SFA findet der Leser in Kapitel 2.

Nachdem die optimalen Funktionen für die gegebenen Trainingsdaten gelernt worden sind, kann man sie auf eine ähnliche Weise untersuchen wie Neurone im sensorischen Nervensystem. Zum Beispiel kann man versuchen, den *optimalen Stimulus* zu bestimmen, das heißt, den Stimulus, der das Ausgangssignal maximiert. Zudem lässt sich das

Antwortverhalten der Funktionen visualisieren, indem man das Ausgangssignal in Abhängigkeit ausgewählter Stimulusparameter aufträgt.

Inhalt der Arbeit

Struktur der Arbeit

Die vorliegende Arbeit untersucht das Langsamkeitsprinzip unter zwei Gesichtspunkten. Im ersten Teil der Arbeit liegt der Schwerpunkt auf der mathematischen Analyse von SFA. Ziel dieses Ansatzes ist zum einen die Entwicklung von Methoden, die analytische Vorhersagen für konkrete Anwendungen von SFA ermöglichen. Zum anderen eröffnet die Analyse ein tieferes Verständnis für den Einfluss der Statistik der Eingangsdaten auf die erlernten Repräsentationen. Der zweite Teil der Arbeit widmet sich der Frage, ob und wie das Langsamkeitsprinzip auf biologisch plausible Weise umgesetzt werden könnte und wie sich die Entwicklung des Antwortverhaltens sensorischer Systeme dynamisch beschreiben lässt.

Teil I: Mathematische Analyse der Slow Feature Analysis

Im ersten Teil der Arbeit wird zunächst gezeigt, dass sich das Optimierungsproblem von SFA unter bestimmten Bedingungen auf partielle Differentialgleichungen abbilden lässt (Kapitel 3 und 5). Diese Gleichungen haben die Struktur von Wellengleichungen und sind eng verwandt mit kanonischen Systemen in der Physik. Diese Analogie ermöglicht den Transfer von Intuitionen und Lösungsverfahren aus der theoretischen Physik.

Im ersten Abschnitt von Kapitel 4 wird die in Kapitel 3 dargestellte Theorie verwendet, um eine neue Methode zur nichtlinearen blinden Quellentrennung zu entwickeln. Die Aufgabe bei der blinden Quellentrennung ist, aus einer unbekannten Mischung unbekannter Signale diese Signale, meist *Quellen* genannt, zu rekonstruieren. Für lineare Mischungen ist dieses Problem im Wesentlichen das der *Unabhängigen Komponenten Analyse* (Independent Component Analysis, ICA; Hyvärinen et al., 2001), da meist angenommen wird, dass die Quellen statistisch unabhängig sind. Im nichtlinearen Fall ist das Problem erheblich komplizierter, da die Annahme der statistischen Unabhängigkeit die Lösung nicht hinreichend bestimmt. Deshalb sind zusätzliche Bedingungen notwendig, um eine eindeutige Lösung des Problems zu erhalten. Langsamkeit ist kürzlich als Kriterium vorgeschlagen worden, um aus der Vielzahl statistisch unabhängiger Lösungen die richtigen auszuwählen (Blaschke et al., 2007). Basierend auf der Theorie aus Kapitel 3 stellen wir einen alternativen Ansatz vor, der bei der Anwendung auf nichtlineare Mischungen von zwei Musikstücken die Quellen in 90% der untersuchten Fälle rekonstruieren kann. Im Vergleich mit dem von Blaschke, Zito und Wiskott vorgeschlagenen ISFA (Independent Slow Feature Analysis) Algorithmus stellt dies einen Fortschritt dar, zumal der neue Algorithmus im Gegensatz zu ISFA keinen Abwägungsparameter enthält, der eine Feinabstimmung erfordert.

Eine weitere Anwendung der Theorie wird in Abschnitt 4.2 vorgestellt. Dort wird gezeigt, dass mit Hilfe von SFA die Position und die Kopfrichtung einer simulierten Ratte aus den visuellen Eindrücken der Ratte in einer virtuellen Umgebung extrahiert werden können. Die Theorie erlaubt hierfür analytische Vorhersagen, die beim Vergleich mit Simulationen eine gute Übereinstimmung zeigen. Die mit SFA erlernbare Repräsentation der Position hat Ähnlichkeit mit einer Repräsentation, die kürzlich im entorhinalen

Kortex von Ratten entdeckt wurde, so genannten Gitterzellen (grid cells; Hafting et al., 2005). Zudem lassen sie sich mit Hilfe von spärlicher Kodierung (*sparse coding*), einem anderem Lernprinzip, in Repräsentationen überführen, die der von Orts- und Kopfrichtungszellen in der hippokampalen Formation von Ratten stark ähnelt. Eine wesentliche Leistung des Modells ist, dass die Repräsentationen direkt aus komplexen und hochdimensionalen visuellen Daten gelernt werden. Dies erfordert die Bildung von komplizierten Invarianzen, da sich zum Beispiel die visuellen Eindrücke beim Drehen des Kopfes stark verändern, obwohl die Position der Ratte unverändert bleibt. Dieses Forschungsprojekt ist eine Kooperation mit Mathias Franzius, der die Simulationen durchführte. Die vorgestellten Ergebnisse wurden veröffentlicht (Franzius, Sprekeler & Wiskott, 2007).

Durch die Arbeit von Berkes und Wiskott (2005) ist bekannt, dass SFA mit quadratischen Funktionen eine Vielzahl von Eigenschaften komplexer Zellen im primären visuellen Kortex reproduzieren kann, wenn als Trainingsdaten quasi-natürliche Bildsequenzen verwendet werden. Ein interessanter Aspekt dieser Simulationen ist, dass die erlernten Eigenschaften nicht von höheren Ordnungen der räumlichen Trainingsbildstatistik abhängen, sondern primär von der Art der Bewegungen, die in den Bildern auftreten. In Kapitel 5 wird ein mathematischer Formalismus entwickelt, der dieses Verhalten auf Invarianzen in der Bildstatistik zurückführt und die analytische Herleitung einiger Simulationsergebnisse erlaubt. Die Theorie zeigt, dass die optimalen Stimuli und die Phaseninvarianz der von SFA erlernten rezeptiven Felder auf Translationen in den Bildsequenzen zurückzuführen sind, wohingegen die Orientierungs- und Frequenzabhängigkeit der Antwort durch Drehungen und Vergrößerungen der Bilder bestimmt werden. Zudem erlaubt die Theorie ein intuitives Verständnis des Verhaltens der simulierten Zellen.

Als Abschluss des ersten Teils der Arbeit wird in Kapitel 6 ein Zusammenhang zwischen den Lernkonzepten der vorhersagenden Kodierung und SFA hergestellt. Dazu wird zunächst eine informationstheoretische Formulierung vorhersagender Kodierung eingeführt. Die anschließende Analyse zeigt, dass diese Formulierung für reversible Eingangsdaten mit Gauss'scher Statistik eng mit SFA verwandt ist. Anschließend werden die Voraussetzungen, die dieser Zusammenhang erfordert, kritisch diskutiert. Die Ergebnisse dieses Kapitels sind in Kooperation mit Felix Creutzig entstanden und werden in Kürze veröffentlicht (Creutzig und Sprekeler, 2008).

Teil II: Zur biologischen Plausibilität des Langsamkeitsprinzips

Die Ergebnisse von Berkes und Wiskott (2005) und von Franzius, Sprekeler & Wiskott (2007) zeigen, dass SFA einige Eigenschaften neuronaler Signalverarbeitung im Nervensystem reproduzieren kann. Dies kann als Indiz dafür gewertet werden, dass das Langsamkeitsprinzip eines der Prinzipien ist, die das Gehirn zum Erlernen seiner Repräsentationen verwendet. Allerdings ist die Implementierung des Langsamkeitsprinzips in Form von SFA biologisch nicht plausibel, insbesondere weil der Algorithmus auf der Lösung eines Eigenwertproblems beruht, einer Aufgabe, die für ein Neuron schwer zu lösen sein könnte.

Aus diesem Grund in Kapitel 7 untersucht, ob das Langsamkeitsprinzip mit Hilfe bekannter synaptischer Plastizitätsmechanismen implementiert werden kann. Zunächst wird ein stetiges Neuronenmodell analysiert. Dabei wird SFA mit einem anderen Lernalgorithmus für das Lernen von Invarianzen, der von Földiàk (1991) eingeführten Spurregel (*trace rule*), verknüpft. Anschließend wird gezeigt, dass sich das Langsamkeitsprinzip in spikenden Neuronen unter Umständen durch einen synaptischen Plastizitätsmechanismus

umsetzen lässt, der von der relativen Zeitabfolge prä- und postsynaptischer Aktionspotentiale abhängt und in der letzten Dekade intensiv untersucht wurde (spike-timing–dependent plasticity, STDP). Die Ergebnisse dieses Kapitels sind in Zusammenarbeit mit Christian Michaelis entstanden und bereits veröffentlicht (Sprekeler et al., 2007).

In Kapitel 7 wird unter anderem gezeigt, dass sich STDP unter bestimmten Bedingungen als Gradientenabstieg auf einer Zielfunktion interpretieren lässt, die langsame Signale bevorzugt. Deshalb widmet sich Kapitel 8 dem Studium der Lerndynamik von gradientenbasiertem Langsamkeitslernen einerseits und und STDP andererseits. Dabei wird besonderes Augenmerk auf die Dynamik von rezeptiven Feldern gelegt. Es wird gezeigt, dass sich diese Dynamik unter bestimmten Bedingungen durch Drift-Diffusions-Gleichungen beschreiben lässt. Zusätzliche Bedingungen an das Aktivitätsschema der Neurone lassen sich in den Formalismus integrieren und ergänzen das dynamische System zu einem Reaktion-Diffusions-System. Diese Klasse von Systemen ist in der Theorie der spontanen Musterbildung weit verbreitet. Die in Kapitel 8 vorgestellten Konzepte stellen einen Ausblick dar, der einen Anknüpfungspunkt zwischen etablierten Methoden der nichtlinearen Dynamik und Modellen rezeptiver Felder herstellt.

Veröffentlichungsliste

Artikel in Fachzeitschriften

- H. Sprekeler and L. Wiskott.
  Analytical Derivation of Complex Cell Properties from the Slowness Principle
  manuscript in preparation

- H. Sprekeler, T. Zito and L. Wiskott.
  Extending Slow Feature Analysis for Nonlinear Blind Source Separation
  submitted

- F. Creutzig and H. Sprekeler (2008).
  Predictive Coding and the Slowness Principle: An Information-Theoretic Approach
  Neural Computation 20(4):1026-41

- M. Franzius, H. Sprekeler and L. Wiskott (2007).
  Slowness and Sparseness lead to Place, Head-Direction and Spatial-View Cells
  PLoS Computational Biology, 3(8):e166

- H. Sprekeler, C. Michaelis and L. Wiskott (2007).
  Slowness: An Objective for Spike-Timing-Dependent Plasticity?
  PLoS Computational Biology 3(6):e112

- G. Kießlich, H. Sprekeler, A. Wacker, and E. Schöll (2004).
  Positive Correlations in Tunneling through coupled Quantum Dots
  Semiconductor Science and Technology 19, S 37

- H. Sprekeler, G. Kießlich, A. Wacker, and E. Schöll (2004).
  Coulomb Effects in Tunneling through a Quantum Dot Stack
  Phys. Rev. B 69, 125328

Tagungsbeiträge

- M. Franzius, H. Sprekeler and L. Wiskott (2007).
  Spike-Timing-Dependent Plasticity and Temporal Input Statistics
  Poster, Computational Neuroscience Meeting (CNS) 2007, Toronto, Kanada.

- H. Sprekeler, C. Michaelis and L. Wiskott (2007).
  Slowness: An Objective for Spike-Timing-Dependent Plasticity?
  Poster, 7th Meeting of the German Neuroscience Society, Göttingen.

- M. Franzius, H. Sprekeler and L. Wiskott (2007).
  Slowness leads to Place Cells and Head Direction Cells
  Poster, 7th Meeting of the German Neuroscience Society, Göttingen.

- H. Sprekeler and L. Wiskott (2006).
  Analytical Derivation of Complex Cell Properties from the Slowness Principle
  Poster, Mathematical Neuroscience Meeting 2006, Sant Julià de Lorià, Andorra.

- M. Franzius, H. Sprekeler and L. Wiskott (2006).
  Slowness leads to Place Cells
  Poster, Computational Neuroscience Meeting (CNS) 2006, Edinburgh, Schottland.

- H. Sprekeler and L. Wiskott (2006).
  Analytical Derivation of Complex Cell Properties from the Slowness Principle
  Poster, Computational Neuroscience Meeting (CNS) 2006, Edinburgh, Schottland.

Berlin, 25. März 2008

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel und Quellen angefertigt habe.

Ich habe mich nicht anderwärts um einen Doktorgrad beworben und besitze derzeit keinen Doktorgrad.

Ich bin in Kenntnis der zugrundeliegenden Promotionsordnung.

Berlin, 25. März 2008